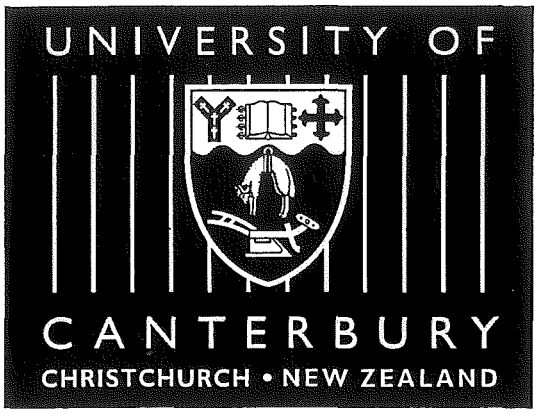
Department of Mathematics and Statistics



Survey Designs and Compensation Methods for Nonresponse Problems

---

A thesis submitted in

partial fulfilment

of the requirements of

the Degree for

PhD in Statistics

at the

University of Canterbury

by

Taweesak Siripornpibul

2001

To ***Poonsak*** and ***Narongdej***

Abstract

This thesis is a comparative study of the unit nonresponse problem. Simulation based on the summary report of the 1997 Thailand Industrial Survey is used to investigate efficient survey designs. Research questions are (1) Which design is best for conducting large-scale surveys such as the Thailand Establishment Survey with nonresponse problems? (2) Which compensation method can reduce the effects from nonresponse problems?

This study begins with an extensive review which brings together the theory of the methods for compensating nonresponse. The simulation compares across all combination of survey designs and compensation methods, thus extending the work of other researchers. In the simulation study there are various sampling designs and compensation methods. Other factors that are considered are sample size, response rates, and types of response (dependent and independent). Nonrespondent subsampling is used to compensate unit nonresponse during the data collection phase. Nonrespondent subsampling methods used in this thesis are one- and two-subsampling schemes. Weighting adjustment procedures and imputation methods are used in the estimation phase. The response mechanisms for weighting adjustment procedures and imputation methods are a naive model and a random homogeneity (RHG) model. In addition a new weighting adjustment method called the bias-removal adjustment method is proposed. Methods for dealing with nonresponse are compared by using bias, variance and design effect.

The main conclusion were, whenever possible, a complex survey design should be used, e.g. stratified or post-stratified sampling with unequal probability of selection. The best method to compensate for nonresponse is nonrespondent subsampling. If subsampling is too costly RHG model with weighting adjustment or with imputation is recommended. For example in weighting adjustment method, the population-based should be used in equal probability sampling with or without replacement. For unequal probability sampling, the sample-based methods should be used in sampling without replacement and the bias-removal method should be used in sampling without replacement. In imputation method, multiple imputation with regression or methods related with regression should be used combined with weighting adjustment procedures described above for each survey design. An algorithm for dealing with nonresponse is presented.

Acknowledgements

I wish to express my sincere appreciation to many people who have helped in this thesis for their support during my time with them.

My first gratitude goes to Dr. Easaw Chacko, my supervisor, for his patient guidance and friendship. Then, I would like to thank Dr. Jennifer Ann Brown, my co-supervisor, for her generous help, advice and support. I also gratefully acknowledge the support I have received from the Department of Mathematics and Statistics at the University of Canterbury.

Special thanks are due to Ms. Jirawan Boonperm, director of Economic Statistics Division, National Statistical Office, Thailand, for her encouragement and suggestion; to Ass.Prof. Watcharaporn On-Seng, Naresuan University, for her encouragement during I study in New Zealand.

For financial support I thank the Ministry of Foreign Affairs and Trade, New Zealand, for NZODA study awards and Naresuan University, Thailand, for support me the last year of study.

I owe my parents and relatives a great deal for their love and support. Among them my cousin, Poonsak and Narongdej, who are not with us anymore. This thesis is dedicated to them.

Finally I thank my mom, Kompor, for her love, willingness and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, I review the history and background of the Thailand establishment surveys in section 1.1. Research objectives are presented in section 1.2. Outline of thesis are presented in section 1.3.

1.1 History and Background

In 1963 the National Statistical Office of Thailand (NSO) was officially established (*National Statistical Office*, 1990) as a department under the Office of the Prime Minister. The NSO is responsible for the statistical system of the country. According to legislation, the NSO is solely in charge of the statistical projects and activities of the country. However, at present, various government agencies and enterprises also conduct a number of statistical projects. These projects are primarily focussed on data collection to serve the management and administrative works of their own organisations. The present statistical system of Thailand is a decentralised system. Nonetheless, the NSO still plays a leading role in compiling fundamental statistics which are directly related to the social and economic situation of the country through a number of significant censuses, and various large scale sample surveys.

The *NSO* has conducted many statistical projects with the purpose of collecting and compiling information on industry, business trade and services, labour force and migration in order to serve the needs of both government and private sectors. The information obtained from these censuses and surveys are used in formulating social and economic development plans, constructing the National Accounts and constructing the Input and Output table. This table is a summary of the national economic activities which are systematically grouped into industrial activities such as agriculture, mining, manufacturing and so on. The private sector also uses such data in making investment decisions.

The Industrial Survey or Census has been conducted periodically by the *NSO* since 1964 (*National Statistical Office*, 1995). The main objective is to collect basic industrial information on the following: number of establishments, number of people engaged, number of employees, compensations, value of raw materials, parts and components purchased, sales values of goods produced for resale, inventory and value of fixed assets.

In the first industrial census in 1964, the United Nation recommendations regarding the concepts, definitions, methods of enumeration, processing and tabulation plans and quality control were adopted, though with some minor modifications. All industrial establishments and household economic units were under the coverage of this census. In municipal areas[1], the complete listing, called the listing frame, was used and complete enumeration was employed. In the non-municipal areas only villages which were occupied by at least two industrial factories (except small rice-mills) were completely enumerated. Data on the type of industry, form of the legal and economic organisation, number of persons engaged and their compensation of employees, cost of production, sales and fixed assets were asked in the census questionnaires. The census data were useful for policy-makers in both the public and

[1] See appendix A

private sector. The data obtained from the census are used as bench-mark data for subsequent surveys. The industrial census was planned to be carried out every 10 years, but because of some limitations such as budget constraint, and insufficient facilities and personnel, the second census was only carried out in 1997.

In addition to the census project the *NSO* has conducted industrial business trade and services surveys more frequently. The annual Industry Survey has been carried out since 1968. It is designed to compile basic data on each type of industry such as number of establishments, employees, value and cost of production and value of sales. The collected data serves the need of policy-makers, researchers and decision-makers in the public and private sectors.

Major surveys of establishments e.g., the Industry Survey and the Survey of Business Trade and Services, originally used a mail questionnaire approach in collecting the data. Questionnaires were sent to the sample establishments according to the address in the listing frame with a request to send them back to the *NSO* within 15 days after receipt. If questionnaires were not sent back a follow-up would be undertaken by telephone or direct visit to the establishment by an enumerator who was a permanent staff working in a provincial office[2]. However, because of the poor response rate, the *NSO* decided from the year 1992 onward to substitute the mail questionnaire approach with a face-to-face interview. For large-scale establishments, as decided by the provincial officer, the survey questionnaires are sent by mail and followed up by personal visits of the enumerator. For the small-scale establishments, enumerators are sent to the sample establishments to conduct the interviews.

The sampling frame for the survey and census of industry is the listing of establishments (*United Nations*, 1994). The interval of the establishment censuses was rather wide and it was necessary to conduct a census of the listing of establishments

[2]See appendix A

in order to provide a sampling frame for establishment surveys during the inter-censal period. The latest establishment listings were carried out in 1984 and 1986. The listing in 1984 covered all establishments in the municipal and sanitary district areas[3] while the listing in 1986 covered those establishments outside the municipal and sanitary district areas. The listings have been used as a sampling frame for almost every subsequent establishment survey. Each year registration data compiled from many sources e.g., the Ministry of Industry, the Ministry of Commerce, the Board of Investment etc, as well as data obtained from the field operation of the other establishment survey projects are used in updating the sampling frame.

Generally, establishment surveys select all listed establishments in the municipal and sanitary district areas, whereas those in the villages are enumerated on a sample basis. The sample design commonly used in the establishment survey is stratified sampling (*National Statistical Office*, 1995). The total area of the country is stratified into two strata by groups of provinces i.e. stratum I comprises the Bangkok Metropolis Area and the other five surrounding provinces and stratum II comprises the remaining 69 provinces. The Bangkok Metropolitan Area and vicinity has been the centre of the economic activities and other activities of Thailand. A large number of industrial and business trade establishments are located in these areas and the pattern and characteristics of such activities are quite distinctive from those located elsewhere in the country.

Data obtained from these surveys are processed by the electronic data processing at the central office. The data are processed according to a data processing package of programs which are designed in advance. The package consists of the preparation of raw data, manual editing and coding, data entry, machine editing and updating, weighting and tabulations. In past surveys it has taken 6 months to conduct all of the processing steps. This is mainly because questionnaires were sent back from the

[3]See appendix A

establishments very late although every attempt had been made to accelerate the return of the questionnaires.

Reports of the survey results are published in two sets, one for the Bangkok Metropolis and vicinity and the other for the whole kingdom. Important characteristics are classified by industry (5-digit code of industry according to the Thailand Standardisation Industrial Code) and establishment size (large and small).

1.2 Research Objectives

The nonresponse rate in the Thailand Establishment Survey is quite high, approximately 70%. The objective of this research was to find optimal methods to reduce errors due to this problem.

There are many articles and textbooks that present methods for dealing with nonresponse. Some examples where methods for dealing with nonresponse are reviewed include *Madow et al* (1983, chapter 4) where 11 case studies using weighting adjustment procedures and imputation methods are presented. *Lessler & Kalsbeek* (1992, pp. 232) summaries 15 case studies that also use weighting adjustment procedures and imputation methods. *Holt & Elliot* (1991) review weighting adjustment procedures. A range of imputation methods are reviewed by *Jinn et al* (1989). Other examples where imputation methods are compared are *Nordholt* (1998), *Schenker & Taylor* (1996). *Sarndal et al* (1992) give the theorem with the nonrespondent subsampling method and with the sample-based adjustment in unequal probability sampling without replacement but not with replacement. However, almost all these compensation methods are considered with equal probability sampling. Moreover there are no comparisons across all these methods. Further there is little guidance for survey planners.

In this thesis I present a case study that compares methods for compensating unit

nonresponse. In addition, the study also compared different survey designs to investigate the interaction between survey designs and compensation methods for dealing with nonresponse. I conclude with an algorithm for dealing with nonresponse.

Methods to reduce error due to nonresponse either involve using a suitable sample design to collect data to get more precise estimates within some constraints or compensating the problem with special statistical techniques. The two main research questions in this study were:

1. Which design is best for conducting large-scale surveys such as the Thailand Establishment Survey with nonresponse problems?

2. Which compensation method can reduce the effects from nonresponse problems?

Given these research questions, the main aims were to compare:

1. Unequal probability sampling designs and equal probability sampling designs both when there is full response and when there is nonresponse.

2. Stratified random sampling, simple random sampling and post-stratified random sampling both when there is full response and when there is nonresponse.

3. Nonrespondent subsampling, weighting adjustment and imputation methods when there is nonresponse.

4. The efficiency among the three above compensation methods in unequal probability sampling design when there are different correlation coefficients between an auxiliary variable and the characteristic of interests.

Another aim was to collect and bring together the range of methods for compensating nonresponse and to present these with their relevant theorems. Proofs for theorems are given where these are not explicitly presented in the literature.

1.3 Outline of the Thesis

This thesis is divided into 7 chapters. Chapter 1 is the general introduction of the thesis. Chapter 2 reviews the sampling process. Three compensation methods for the unit nonresponse problems such as nonrespondent subsampling, weighting adjustment procedures and imputation methods which are described in chapter 3, 4 and 5 respectively. Chapter 6 presents the simulation methods and results of these three compensation methods. Conclusions are presented in chapter 7.

Chapter 2

Review of the Sampling Process

In this chapter I introduce some concepts of sampling techniques which will be used and referred to in the following chapters. In section 2.1, I discuss collection methods for census and sample surveys. Section 2.2 reviews two main categories of surveys: descriptive and analytical surveys. Section 2.3 presents sample designs including sampling plans, estimation procedures and basic sampling designs. Section 2.4 reviews two inference approaches: frequentist and Bayesian approach. Section 2.5 presents sampling and non-sampling errors. Section 2.6 reviews nonresponse problems, methods for controlling and compensating for this problem, mechanisms of nonresponse and inference approaches to non-responded samples. In section 2.7, the design effect and misspecification effect, used to compare efficiency between surveys, is defined.

2.1 Census and Survey

In general, information on the population can be collected in two ways. The first way is to enumerate every unit in the population. This complete enumeration is called a census. In many countries, a national statistical agency such as US Bureau of the

census, Australian Bureau of Statistics and Statistics New Zealand, has authorisation to conduct censuses to provide relevant information, e.g., census of population and housing, industrial census, etc. The second way to collect information is by using a sample survey where the survey itself and possibly the enumeration is limited to a subset of the population. Government organisations and non-government organisations frequently conduct repeated sample surveys to estimate current monthly or quarterly statistics. One-time sample surveys are often conducted to draw inference about a current issue.

There are some relationships between census and sample survey. For example (i) sampling is often used to evaluate the census process. The post enumeration survey (*PES*) (*US Bureau of the census*, 1970) is a special sampling technique to assess the coverage and content errors of the census; (ii) a census can provide some information to make a frame from which sample units can be drawn (*United Nation*, 1982); (iii) information obtained from a census, even if out of date, can be used to provide supplementary information for improving the efficiency of sample designs (*Som* 1996); and (iv) sampling can be used as an integral part of a census as in a partial enumeration census (*National Statistical Office*, 1990). Thus, census and sample survey are complementary and not competitive.

An example of complementary nature of census and sample surveys is the 1990 census of population and housing in Thailand. The census had two forms for interviews of individuals. The first form was used for every unit in the population. The second form, which had more extended questions, was used only for units selected using a systematic scheme. General information was taken from the first and second forms to summarise fundamental issues. The extended questions in the second form were taken to be a survey and used for analysis. The *NSO* also planned to use *PES* to evaluate the coverage and content of error in the 1990 census of population and housing. In some cases, when non-sampling errors occur, the results from *PES*

can be used for adjusting or revising the final results before they are released to the public. As another example, a Labour Force Survey (*National Statistical Office*, 1996) can use information from the census of population and housing as a frame to make decisions about sampling design and estimation procedures.

By comparison with a census, a sample survey is based on collecting data from a small number of units from a population. A sample survey generally requires different resource schedules for designing and executing. A sample survey will usually be less costly in total than a census but the cost per unit of observation may be higher. Surveys may also permit collection of a wider range of data and allow more choice of data collection methods. Furthermore, sampling can generally provide more accuracy than a census with perhaps the exception of small populations. Except in small populations, non-sampling errors in a sample survey are more easily controlled than in a census. However, occasionally it is necessary to take a census in order to get information relevant to maintaining a sampling frame.

2.2 Descriptive and Analytic Surveys

Sample survey can be divided into two board categories based on the aim of the study: *descriptive surveys* and *analytic surveys.*

In descriptive surveys, summary measures of the population, such as means and totals, need to be precisely and efficiently estimated. For example, in industrial surveys, the parameter often estimated is the average number of persons engaged in different occupational groups.

In analytic surveys, the main purpose is the comparison among subgroups of the population. For instance, in Labour Force Surveys, the interest is not only in the average number of hours worked per day and the wages paid but also in whether men work longer hours than women and whether men receive higher wages than

women for the same type of work.

Efficiency of the sample survey is important. Stratification and auxiliary information, e.g., business sizes and types, can be beneficial in sample estimation in order to increase the efficiency of the estimates both in descriptive and in analytic surveys.

Inference in descriptive surveys is concerned exclusively with a fixed population, although super-population and other models are often used in the estimation. In descriptive surveys, target parameters are objectives determined before the data are collected or analysed while for analytic surveys the parameters of interest are not fixed in advance but evolve through an adaptive process as the analysis progresses (*Skinner et al* , 1989).

2.3 Sample Design

In the planning stage of the survey, survey objectives must be clearly defined as commensurate with available resources in terms of money, manpower and the time limit specified for the survey. If survey objectives are not well defined, there can be bias or non-sampling errors in the estimates. Survey objectives should specify the survey variables, methods of observation, methods of analysis, utilisation of survey results and desired precision of survey results. However, practical realities of sample design often influence and change the survey objectives. The sample design and survey objectives are thus related.

The *sample design* consists of the defined sampling plan and estimation procedure. Different sample designs would result in different errors, and choosing the design with the smallest error is the principal aim of sample design.

2.3.1 Sampling Plan

The *sampling plan* refers to what a sample consists of and how the sample is to be obtained. It is a set of specifications which describe the population and the parameters of interests, the sample and statistics, the frame, the sampling units, the sample size and the sample selection methods.

2.3.1.1 Population and Parameters

One major task of a sampling plan is to identify the *population* to be studied consistent with the survey objectives. The population is the aggregate or collection of *elementary units*, sometimes just called *units*, for which information is sought.

The population may be defined in terms of (i) content, (ii) units, (iii) extent and (iv) time. For example, in the design of an establishment survey a researcher may specify: (i) the content as all establishments; (ii) the units as in industrial establishment units; (iii) the extent as in Thailand; and (iv) the time as 1999. This defined population is what ideally is to be studied and is called the *target population.* However, often the population actually used must be refined to obtain a practicable *survey population* or *coverage.* For example, the above might be refined as: (i) all establishment with 10 or more persons engaged; (ii) in industrial establishment units; (iii) in Thailand; and (iv) during the period January 1 to December 31, 1999. The researcher should be aware of any gaps between the target and survey population and understand that the conclusions only apply to the survey population.

Survey objectives can often lead to the population being subdivided into groups, called sub-populations, in the planning stage of the survey. Sometimes the population cannot be subdivided at this stage but can be subdivided after collecting data into groups called domains of study. These groups can then be used for separate estimates in the analysis phase. Since the results are obtained separately for each

domain of study, any error in the specification of the domain may lead to difficulties in the interpretation of the survey results. Depending on the question of interest, the domain of study may also be the target population.

Consider a finite survey population $\mathbf{U} = \{u_1, ..., u_k, ..., u_N\}$ of N elements labelled from 1 to N. Let Y denote the characteristics of interest called the *study variable*, with unknown population value $y_1, ..., y_k, ..., y_N$. In some case an auxiliary variable X and an indicator variable, $\mathbf{I} = I_1, ..., I_k, ..., I_N$, are also used. The population values of X, which is usually assumed known for all the N population elements, are denoted by $x_1, ..., x_k, ..., x_N$. The indicators I_k equals 1 if the k^{th} unit is chosen in a sample, otherwise it is 0.

The ultimate objective of any sample survey is to make inferences about a population of interest. Such inferences are based on information contained in a sample selected from that population. The researcher usually aims at the estimation of certain unknown features of the population. These population characteristics are called *population parameters* or simply *parameters*. Any real valued function of $y_1, ..., y_N$ is known as a parameter.

Typical parameters are the total, the mean and the variance. They are defined as follows:

$$\begin{aligned} \text{Total } \tau &= \sum_{k=1}^{N} y_k = y_1 + ... + y_k + ... + y_N, \\ \text{Mean } \mu &= \tau/N, \\ \text{Variance } \sigma^2 &= \frac{1}{N}\sum_{k=1}^{N}(y_k - \mu)^2, \\ \text{or } S^2 &= \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu)^2. \end{aligned}$$

2.3.1.2 Sample and Statistics

A subset of the population selected for data collection is called a *sample*. Let $\mathbf{S} = \{S_1, ..., S_k, ..., S_n\}$ or { i selected in the sample } denote a sample and the values

of the variable in the sample are denoted by $y_1, ..., y_k, ..., y_n$. Since a sample cannot give the true parameter value unless $n = N$ and sampling is without replacement, the parameter have to approximated by using an *estimator*. An estimator of the population parameter is a specific computational formula or algorithm which is used to calculate an approximation called a *statistic* to the desired parameter from the selected sample.

The following three estimators are often used for *equal probability sampling* (defined later in sample selection method subsection):

$$\begin{aligned} \text{Mean} \quad \hat{\mu} &= \frac{1}{n}\sum_{k=1}^{n} y_k, \\ \text{Total} \quad \hat{\tau} &= N\hat{\mu}, \\ \text{Variance} \quad s^2 &= \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \hat{\mu})^2, \end{aligned}$$

where n is the sample size and $y_k = y_i$ if the k^{th} sample chooses the i^{th} element.

2.3.1.3 Frame

To facilitate a sample to be drawn from the population the units are organised into what is called a *sampling frame*. The sampling frame, or often simply called the *frame*, is the set of all sampling units organised in the form of either a list of single elementary units or a group of elementary units.

The frame may lead to *single-stage* or *multi-stage* survey designs (*Som*, 1996 or *Foreman*, 1991). If each elementary unit can be sampled directly, it is called a single-stage survey design and the corresponding frame is called as a *listing frame*. When it is not possible to sample elementary units directly, elementary units are grouped with known common properties and the resulting frame is called an *enumeration frame* and this leads to a multi-stage survey design. At each stage of a multi-stage survey design, each of the groups of units chosen has a sub-frame constructed and at the final stage the sub-frame is a listing frame. For example, in the

yearly Labour Force Survey in Thailand, a household listing was not available at the planning stage of this survey and to construct a complete listing frame within a short period would have been expensive. A stratified two-stage sampling was adopted for use in a certain province. In the first stage, the frame consisted of groups of households such as a village in a rural area or a block on a map in an urban area. At the second stage a listing frame was constructed from the blocks or villages that were sampled (*National Statistical Office*, 1996).

2.3.1.4 Sampling Units

The sampling frame divides the population into a finite number of distinct, non-overlapping and identifiable units called *sampling* or *survey* units, so that each member of the population belongs to only one sampling unit. The type of sampling unit depends on the nature of the study. For instance, a business establishment may be considered the sampling unit in an industrial survey; a farm or a group of farms owned or operated by a household in a crop survey. The sampling unit, whether elementary units themselves or groups of elementary units, should be well defined.

Sampling units may or may not correspond to the units of analysis. An example where the sampling unit may be different from the unit of analysis is the case of a household survey. The unit selected may be a dwelling, whereas the units of analysis may be individuals or families within the dwelling.

2.3.1.5 Sample Size

One of the first consideration in the planning of a sample survey is the size of the sample. The sample size is the number of distinct elementary units in a sample S.

In any decision related to the precision expected of the sample survey, a number of factors must be taken into account. Generally, the factors which decide the scale

of the survey operations are cost, time, operational constraints and the desired precision of the results. Such properties as population size, variability of characteristics in the population and sample plan will all affect the precision of the estimates. Consequently, all these factors have to be considered in the statistical formulae which ultimately relate sample size to the desired level of precision.

An increase in sample size will lead to an increase in the precision of estimates of the parameters of interest. However, the sampling cost will also typically increase. Thus, there has to be a trade off between cost and precision in desired sample size.

2.3.1.6 Sample Selection Methods

The method which is used to select the sample from a population is known as *sample selection method* or *sampling procedure.* Its aim is to obtain the estimates of the population characteristics of interest and their precision. The procedure can be distinguished by choices from (i) probability or non-probability sampling, (ii) equal or unequal probability sampling and (iii) sampling with or without replacement.

i) Probability and Non-probability Sampling

There are two types of sampling methods based on the probability of selection of the elements: *probability sampling* and *non-probability sampling.*

Probability sampling method is a method whereby every element in the population has a known nonzero probability of being included in the sample, i.e. $p(\mathrm{S}) > 0$ for all S. It is desirable because it eliminates bias in estimation of the parameter. A good sample will be as free from selection bias as possible. Selection bias occurs when some part of the target population is not in the sampled population. A random sample is sometimes also called a *probability sample.* Random samples can produce valid estimates and measures of reliability of estimates called *sampling errors.* It also enables the theoretical values of the variance estimates of parameter to be computed.

Probability sampling is often a time consuming and expensive procedure and may not be feasible in many situations, such as in remote areas or hill tribes, and it may be necessary to choose a sample by using *non-probability sampling* also called *non-random sampling* methods. Non-probability sampling is quite frequently used especially in market research and public opinion surveys. In this sampling method, the degree of reliability of the sample results cannot be measured. An example of non-probability sampling is where inexpensive information is collected by asking persons known to be experts in the subject. This is called *judgement sampling.* Alternatively a fixed numbers of individual from certain demographic subpopulations as sex or race are interviewed. The specific selection is often left in the hands of the interviewers. This method is called *quota sampling* and tends to be highly biased. Another type of non-probability sampling is *convenience sampling* also often called *haphazard sampling.* These arise when samples are made up of individuals causally or conveniently available, e.g., the customers passing through a checkout or passengers at a door gate.

ii) Equal and Unequal Probability Sampling

Probability sampling methods are general divided into two types: *equal probability sampling* and *unequal probability sampling.*

If every elementary unit in the population has the same chance of being chosen in the sample, this sampling method is called equal probability sampling (*EPSEM*). Equal probability sampling provides schemes that are simple to design and to analyse. Such schemes are therefore popular and their range of applicability is wide and varied (*Satin & Shastry*, 1993).

On occasions, however, the sampling frame can provide some useful quantitative information. If the value of this information is closely related to the study variable and known for all the population units, it could be utilised in selecting the sample to increase the efficiency of estimators. This procedure for selecting units into the

sample is known as unequal probability sampling or *varying probability sampling.* There are many schemes that use unequal probability sampling. The most common type of unequal probability sampling is sampling with probability proportional to size related to the value of the auxiliary variable (*Sarndal et al*, 1992).

iii) Sampling with and without Replacement

If a unit can occur only once in a sample, it is called sampling without replacement; otherwise it is called sampling with replacement. In both cases if a sample of size n is drawn from a population of size N in such a way that every possible sample of size n has the same chance of being selected, i.e. chosen with equal probability, it is called *simple random sampling (SRS).*

- *Simple random sampling without replacement (SRSWOR)*: If a sample of size n is drawn from a population of size N in such a way that $\binom{N}{n}$ possible samples of n elements has the same probability of selection, $1/\binom{N}{n}$. The probability of any one element being selected is equal to $\frac{n}{N}$. The selection of subsequent elements is dependent on previous selection.

- *Simple random sampling with replacement (SRSWR)*: If a sample of size n is drawn from a population of size N in such a way that N^n possible samples of n elements has the same probability of selection, $(\frac{1}{N})^n$. The probability of any one element being selected is equal to $\frac{1}{N}$. All selections are independent since the selected unit is restored to the population before making the next selection.

Generally, sampling with replacement is wasteful and does not serve a useful purpose. Moreover, for a given sample size n, sampling without replacement has less variance than sampling with replacement. However, sampling with replacement is an attractive method to use with more complex sampling design such as unequal

probability sampling because of the simplicity with which its exact variance can be estimated.

In general the sample selection is dependent with the previous sampling unit chosen in $SRSWOR$. However, if the sample size is very small compared to the population size (i.e. $N \gg n$), the sample selection mechanism is often assumed independent. Thus, the estimation procedure for sampling without replacement approximates closely that under the independent identically distributed (IID) approach for sampling with replacement, simplifying the analysis. Nowadays many statisticians use $SRSWR$ as a standard for comparing other designs (*Kish*, 1995).

2.3.2 Estimation Procedure

The basis of estimation procedures is the sampling weight given to the unit. The sampling weight for the unit tends to be the inverse of the probability of selection of the unit in the sample. It indicates the number of units in the population that are represented by a unit in the sample. For example, in simple random sampling design, the estimated total of a study variable Y is

$$\hat{\tau}_{srs} = \frac{N}{n} \sum_{k=1}^{n} y_k,$$

where $W = \frac{N}{n}$ is the sampling weight since the probability of selection of each unit is $\frac{n}{N}$.

The use of sampling weights to produce estimates of population characteristics will be affected by some techniques for handling the problem of information which may be completely or partially unavailable for some units in the sample. For instance, if complete nonresponse had occurred in the sampling design example above, two factors must be taken into account in calculating weights for units selected: (i) the basic selection probability for each unit selected; (ii) the nonresponse factor ap-

plied to each responding unit to compensate for units for which there was a complete nonresponse to the survey. The combination of these two factors is then

$$W^* = (\frac{N}{n})(\frac{n}{m}) = \frac{N}{m},$$

where m is the number of responding units, giving the estimated total of Y as

$$\hat{\tau}_{srs} = \frac{N}{m} \sum_{k=1}^{m} y_k.$$

However, the most common question researchers ask themselves is how to ensure good estimates from data. The researcher has to decide on the type of estimator based on the properties of estimators.

2.3.2.1 Types of Estimator

In general, there are two broad categories of estimators for the population characteristics. A *simple estimator* is a procedure which uses only the study variable. The estimators in the section above are examples of simple estimators. A *composite estimator* uses a combination of the study variable and the auxiliary variable to increase the precision of the estimate.

Composite estimators can be divided into two different types: *ratio estimator* and *regression estimator*. In *ratio estimators*, the weights of the units of the population are adjusted by a multiplying factor. This factor is the ratio of the external data value (an auxiliary variable obtained not from the sample) and the sample estimate. It is important that the external data source pertain to the same population and be based upon comparable concepts, definitions, reference periods, etc. as that of the survey. For example, in SRS design, if an auxiliary variable X has been used to increase the efficiency, then, a ratio estimator of the total estimate of Y is given by

$$\hat{\tau}_r^{srs} = \frac{\hat{\mu}_y}{\hat{\mu}_x} \tau_x = \frac{\sum_{k=1}^{n} y_k}{\sum_{k=1}^{n} x_k} \tau_x,$$

where τ_x is the total population amount of the auxiliary variable X.

In some sampling situations, there may be an auxiliary variable X which is linearly related to the study variable Y, at least approximately and without a zero intercept. In this situation, a *linear regression estimator*, rather than a ratio estimator, might be appropriate. For instance, in SRS scheme a regression estimator of the population total τ_y is

$$\hat{\tau}_l^{srs} = N\hat{\mu}_y + b(\tau_x - N\hat{\mu}_x),$$

where $b = \frac{\sum_{k=1}^{n}(x_k-\hat{\mu}_x)(y_k-\hat{\mu}_y)}{\sum_{k=1}^{n}(x_k-\hat{\mu}_x)^2}$ *and* $\tau_x = \sum_{k=1}^{N} x_k$.

In a special case of regression estimation, b is taken to be 1, and in this case, the estimated total of τ_y is $N\hat{\mu}_y + N(\mu_x - \hat{\mu}_x)$ and this is called a *difference* estimator.

2.3.2.2 Properties of Estimator

The challenge in estimation is finding "good" estimates of the population parameters. For example, from elementary statistics, the sample mean, $\hat{\mu}$, is used to estimate the population mean, μ. Alternatively the sample median, $\tilde{y}$ could be used to estimate μ. The question is which of these better? To determine the answer, some criteria must be used to find out what is meant by "good".

A good estimator is generally considered to have the properties of: (i) unbiasedness, (ii) consistency, and (iii) efficiency

i) *Unbiasedness*: An estimator $\hat{\theta}$ of θ is said to be an unbiased estimator of θ if and only if $E(\hat{\theta}) = \theta$. If $\hat{\theta}$ is not an unbiased estimator for θ, the bias of $\hat{\theta}$ is given by

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

ii) *Consistency*: An estimator $\hat{\theta}$ is said to be a consistent estimator of θ if the probability of the estimator exceeding θ in absolute value by any given small amount, ε, tends to zero as the sample sizes n tends to population size

$$\lim_{n \to N} P(|\hat{\theta} - \theta| < \varepsilon) \to 0.$$

A consistent estimator is not necessarily unbiased. If, however, it is biased, the bias will tend to zero in the limit as the sample size tends to population size.

The value of $\hat{\theta}$ obtained from a given sample will generally be different from θ. The difference, $(\hat{\theta} - \theta)$, is the error in the estimation of θ. Let $L(\hat{\theta}, \theta)$ be the loss that will be incurred through an error in the estimation of θ. The expected value of the loss function is called the expected loss or the risk function. The most commonly used loss function is the squared error, namely,

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2.$$

For this loss function, the expected loss is known as the mean square error and given by

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= (\hat{\theta} - \theta)^2 \\ &= var(\hat{\theta}) + Bias(\hat{\theta})^2. \end{aligned}$$

If an estimator $\hat{\theta}$ is an unbiased estimator, then MSE is equal to the variance.

iii) *Efficiency*: Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of a parameter θ. Then, $\hat{\theta}_1$ is said to be more efficient than $\hat{\theta}_2$ if and only if $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$.

In addition to the above, a "good" estimator should also have the properties of reliability, validity and accuracy.

The *reliability* of an estimator refers to how close the estimator is to its expected value over repetitions of the sampling process. If there is no measurement error in

the survey, then the reliability of an estimator is related to the standard error. The smaller the standard error of an estimator, the greater its reliability.

The *validity* of an estimated population characteristic refers to how the mean of the estimator differs from the true value of the parameter being estimated. If there is no measurement error, the validity of an estimator is related to the bias of the estimator. The smaller the bias, the greater its validity.

The *accuracy* of an estimator refers to how far away a particular value of the estimate is, on average, from the true value of the parameter being measured. The accuracy of an estimator is generally evaluated on the basis of its MSE, or on the basis of the square root of its MSE ($RMSE$). The smaller the MSE of an estimate, the greater its accuracy.

2.3.3 Basic Sampling Design

The aim of a sample survey is to find an estimator (mentioned above in section 2.3.2) with a desired precision within a budget by making an inference about the population from information contained in a sample. There are two factors that affect the quantity of information contained in the sample and hence the precision of any inference-making procedure: variation in population and sample size.

The first factor is the amount of variation in population. This variation can frequently be controlled by sample selection. Thus the procedure for selecting the sample is called the *sampling design* or *sample survey design*.

The second factor is the size of the sample. Surveys can be expensive and it is a matter of finding a survey design that minimises the sampling cost subject to achieving the desired precision. Cost per observation can vary according to the sampling design. As this cost per observation depends on individuals circumstances cost is assumed to be dependent on the sample size and fixed cost, $C = C_o + nC_p$,

where C_p is cost per unit sampled and may depend on the unit and on the design chosen and C_o is a fixed cost. For a given budget that fixes the sample size n, the design must be sought with the most precise estimators for this n.

There are eleven basic sampling designs presented in this section. The definition of the sample variance, s^2, the population variance, S^2 or σ^2 as given in section 2.3.1 is used in theorems 2.1-2.6. For convenience, when discussing equal probability sampling, the estimator of the mean is considered and when unequal probability sampling is discussed, the estimator of the total will be considered. The following theorems (2.1 to 2.11) are stated without proof. Proofs can be found in standard sampling textbooks and journals such as *Cochran* (1963, 1977), *Deming* (1950), *Hansen et al* (1953), *Kish* (1965), *Murthy* (1953), *Raj* (1968), *Sukhatme et al* (1984), *Koijin* (1986), *Yates* (1981), *Thompson* (1992), *Sarndal et al* (1992), *Cassel et al* (1977), *Chaudhuri et al* (1988, 1992), *Jessen* (1978), *Scheaffer et al* (1996), *Smith* (1976), *Rao & Bellhouse* (1990), *Sedransk & Smith* (1988), *Thomsen & Tesfu* (1988) and *Rao* (1985).

Three cases of equal probability sampling are considered: (a) Simple random sampling(SRS) with and without replacement, (b) Stratified random sampling (ST) and Post-stratified random sampling (PT) with and without replacement. Unequal probability sampling is also considered under two cases of: (c) RS with and without replacement and ST with and without replacement as well as (d) an approximation for unequal probability selection in without replacement schemes.

a) *SRS with/without replacement*

The most common sampling method is SRS with and without replacement as discussed in subsection (iii) in 2.3.1.6.

Theorem 2.1 *Simple Random Sampling with Replacement (SRSWR)*

In simple random sampling with replacement, an unbiased estimator of μ is

$$\hat{\mu}_{srs} = \frac{1}{n}\sum_{k=1}^{n} y_k,$$

with a variance of

$$V(\hat{\mu}_{srs}) = (\frac{N-1}{N})\frac{S^2}{n} = \frac{\sigma^2}{n}.$$

An unbiased estimator for the variance of the sample mean is

$$v(\hat{\mu}_{srs}) = \frac{s^2}{n}.$$

■

Theorem 2.2 *Simple Random Sampling without Replacement (SRSWOR)*

In simple random sampling without replacement, an unbiased estimator of μ is

$$\hat{\mu}_{srs} = \frac{1}{n}\sum_{k=1}^{n} y_k,$$

with a variance of

$$V(\hat{\mu}_{srs}) = (\frac{N-n}{N})\frac{S^2}{n}.$$

An unbiased estimator for the variance of the sample mean is

$$v(\hat{\mu}_{srs}) = (\frac{N-n}{N})\frac{s^2}{n}.$$

■

b) *ST and PT with and without replacement*

Stratified random sampling is a way of maximising the amount of information for a given cost. If qualitative information is provided within a sampling frame, stratified random sampling may increase the precision of the estimator over simple random sampling for a given sample size. However, if the sampling frame cannot

give information about a key variable to differentiate sample units into strata but units can be classified after data collected, simple random sampling can be used in the planning stage. In the analysis phase stratified random sampling, with a minor modification, is used and this procedure is called post-stratified sampling.

The following notations will be used in Stratified Random Sampling and Post-stratified Random Sampling:

$$\begin{aligned}
\text{Population mean in stratum } h \quad \mu_h &= \frac{1}{N_h}\sum_{k=1}^{N_h} y_{hk}, \\
\text{Sample mean in stratum } h \quad \hat{\mu}_h &= \frac{1}{n_h}\sum_{k=1}^{n_h} y_{hk}, \\
\text{Population variance in stratum } h \quad \sigma_h^2 &= \frac{1}{N_h}\sum_{k=1}^{N_h}(y_{hk}-\mu_h)^2, \\
\text{or} \quad S_h^2 &= \frac{1}{N_h-1}\sum_{k=1}^{N_h}(y_{hk}-\mu_h)^2, \\
\text{Sample variance in stratum } h \quad s_h^2 &= \frac{1}{n_h-1}\sum_{k=1}^{n_h}(y_{hk}-\hat{\mu}_h)^2,
\end{aligned}$$

where N_h and n_h are the population and sample sizes in stratum h respectively.

- *ST with/without replacement*

  A stratified random sample is one obtained by separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum. This may be done with or without replacement.

Theorem 2.3 *Stratified Random Sampling with Replacement (STWR)*

In stratified random sampling with replacement, an unbiased estimator of μ is

$$\hat{\mu}_{st} = \textstyle\sum_{h=1}^{H} \frac{N_h}{N}\hat{\mu}_h,$$

with a variance of

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2(\frac{\sigma_h^2}{n_h}).$$

An unbiased estimator for the variance of the sample mean is

$$v(\hat{\mu}_{st}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2(\frac{s_h^2}{n_h}).$$

■

Theorem 2.4 *Stratified Random Sampling without Replacement (STWOR)*

In stratified random sampling without replacement, an unbiased estimator of μ is

$$\hat{\mu}_{st} = \sum_{h=1}^{H} \frac{N_h}{N}\hat{\mu}_h,$$

with a variance of

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h n_h} S_h^2.$$

An unbiased estimator for the variance of the sample mean is

$$v(\hat{\mu}_{st}) = \sum_{h=1}^{H}(\frac{N_h}{N})^2 \frac{N_h - n_h}{N_h n_h} s_h^2.$$

■

The above two theorems show how the variance is affected by the sample size in each stratum, n_h. The objective of a sample design is to provide an estimator with an acceptable variance at the lowest possible cost. Given a total sample of size n, there are many ways to divide the n sample units into non-overlapping stratum with size $n_1, n_2, ..., n_H$. Each division may result in a different sample variance and different costs. The best allocation scheme is affected by three factors: (i) the total number of elements in each stratum; (ii) the variability of observations within each stratum; and (iii) the cost of obtaining an observation from each stratum. Given a total sample size n, a researcher has to choose a number of stratum (H) and how to allocate it among the H strata.

First assume the same population size in each stratum, the same cost per sample unit and the same population variance for all strata. In this case, an equal sample

size might reasonably be assigned for every stratum, i.e. assuming (i), (ii) and (iii) are the same for all strata, the sample size for stratum h would be

$$n_h = \frac{n}{H}.$$

If N_h are known and differ from stratum to stratum, then *proportional allocation* could be used to maintain a constant sampling fraction throughout the population, i.e. when (i) is known and (ii) and (iii) are assumed the same for all strata, the sample size allocated to stratum h would be

$$n_h = (\frac{N_h}{N})n.$$

If the population size and variance vary from stratum to stratum and assuming the same cost for all strata, the allocation scheme which estimates the population total (or mean) with the lowest variance for a fixed total sample size n under stratified random sampling is *Neyman allocation*, given by

$$n_h = \frac{nN_h\sigma_h}{\sum_{i=1}^{H} N_i\sigma_i} \text{ or } \frac{nN_hS_h}{\sum_{i=1}^{H} N_iS_i}.$$

If the population standard deviation σ_h (or S_h) in a stratum h is not available, the estimate of the sample standard deviation, for example from the past surveys, is often used.

Although this thesis does not consider costs, we briefly look at how n_h would be affected if costs are considered. If the cost of sampling, which is measured in term of time or money, differs from stratum to stratum and the total cost C can be described by the linear relationship

$$C = c_0 + c_1n_1 + ... + c_Hn_H,$$

where C is the total cost of the survey, c_0 is a fixed cost, and c_h is the cost per unit for observing in stratum h, then for a fixed total cost C, the lowest variance is achieved when

$$n_h = \frac{(C-C_0)N_h\sigma_h/\sqrt{c_h}}{\sum_{i=1}^{H} N_i\sigma_i/\sqrt{c_i}}.$$

- *PT with/without replacement*

  Occasionally, sampling problems arise in which a researcher would like to stratify on a key variable which is only available after the sample is observed. Suppose a simple random sample of size n is selected and then divided into n_h for $h = 1, ..., H_s$ after the sample is collected based on an auxiliary information in the sample units. In this situation the n_h and H_s are random since they can change from sample to sample even though n is fixed. This method of stratification is called *post-stratification* or *stratification after the selection of a sample.* Thus, post-stratification is more relevant at the analysis phase. However analysis is only possible when N_h are known or can be approximated. More discussions with post-stratification are in *Little* (1993), *Holt & Smith* (1979), *Jagers et al* (1985), *Jagers* (1986) and *Valliant* (1993).

  Since the n_h are random, a general expression for $V(\hat{\mu}_{st})$ can be approximated by replacing $\frac{1}{n_h}$ by its expected value. The expected value of the reciprocal of a random variable usually has to be approximated and a good approximation as given by *Hansen, Hurwitz & Madow*(1953), is

$$E(\frac{1}{n_h}) \approx \frac{1}{nW_h} + \frac{(1-f)(1-W_h)}{n^2W_h^2},$$

  where $W_h = \frac{N_h}{N}$ and $f = \frac{n}{N}$.

  Using this approximation we have the following two theorems for PT.

Theorem 2.5 *Post-stratified Random Sampling with Replacement (PTWR)*

In post-stratified random sampling with replacement, an unbiased estimator of μ is

$$\hat{\mu}_{pt} = \sum_{h=1}^{H} W_h\hat{\mu}_h,$$

with a variance of

$$V(\hat{\mu}_{pt}) \approx \frac{1}{n}\sum_{h=1}^{H_s} W_h\sigma_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H_s}(1-W_h)\sigma_h^2.$$

An unbiased estimator for the variance of the sample mean is

$$V(\hat{\mu}_{pt}) \approx \frac{1}{n}\sum_{h=1}^{H_s} W_h s_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H_s}(1-W_h)s_h^2.$$

■

Theorem 2.6 *Post-stratified Random Sampling without Replacement (PTWOR)*

In post-stratified random sampling without replacement, an unbiased estimator of μ is

$$\hat{\mu}_{pt} = \sum_{h=1}^{H} W_h\hat{\mu}_h,$$

with a variance of

$$V(\hat{\mu}_{pt}) \approx \frac{1-f}{n}\sum_{h=1}^{H_s} W_h S_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H_s}(1-W_h)S_h^2.$$

An unbiased estimator for the variance of the sample mean is

$$v(\hat{\mu}_{pt}) \approx \frac{1-f}{n}\sum_{h=1}^{H_s} W_h s_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{H_s}(1-W_h)s_h^2.$$

■

In those cases when $\frac{N_h}{N}$ is known and $n_h \geq 20$ for each stratum, *Cochran* (1977) has shown that post-stratification is nearly as accurate as stratified random sampling with proportional allocation.

c) *RS and ST with and without replacement in unequal probability selection*

Unequal probability sampling is an alternative choice to maximise the amount of information for a given cost. If quantitative information is provided within a

sampling frame, unequal probability sampling may increase the precision of the estimator as compared with equal probability sampling for a given sample size. With unequal probability selection, different units in the population have different probabilities of being included in the sample. Thus, a unit which has a higher probability of selection is expected to contribute more to the population total than one with a lower probability selection. Often units with the value of the auxiliary information is high are assigned a high probability of selection. This would lead to a bias in, e.g., the estimator of the total as given by theorem 2.7 to 2.10. In order to overcome the bias, the sample observations are weighted. More details with unequal probability sampling are in *Wolter* (1985), *Sunter* (1986), *Godambe & Thompson* (1988), *Basu* (1971), *Horvitz & Thompson* (1952), *Sarndal* (1980, 1996), *Sarndal et al* (1989) and *Kish* (1992).

- *Random sampling with/without replacement in unequal probability selection*

  The most common way to select a sample with unequal probability selection is random sampling.

Theorem 2.7 *Random Unequal Probability Sampling with Replacement*

In random sampling with replacement and unequal probability selection, an unbiased estimator of τ is

$$\hat{\tau}_{pps}^{srs} = \frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{p_k},$$

where $p_k = x_k / \sum_{i=1}^{n} x_i$ and x_k is a value of X in unit k highly related with y_k, with a variance of

$$V_{srs}(\hat{\tau}_{pps}) = \frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k} - \tau^2].$$

An unbiased estimator for the variance of the sample total is

$$v_{srs}(\hat{\tau}_{pps}) = \frac{1}{n(n-1)}[\sum_{k=1}^{n} \frac{y_k^2}{p_k^2} - n\hat{\tau}_{pps}^2].$$

Theorem 2.8 *Random Unequal Probability Sampling without Replacement*

In random sampling without replacement and unequal probability selection, an unbiased estimator of τ is

$$\hat{\tau}_{\pi ps}^{srs} = \sum_{k=1}^{n} \frac{y_k}{\pi_k},$$

with a variance of

$$V_{srs}(\hat{\tau}_{\pi ps}) = \sum_{k=1}^{n}(\frac{1-\pi_k}{\pi_k})y_k^2 + \sum_{k=1}^{N}\sum_{i\neq k}^{N}(\frac{\pi_{ki}-\pi_k\pi_i}{\pi_k\pi_i})y_k y_i.$$

An unbiased estimator for the variance of the sample total is

$$v_{srs}(\hat{\tau}_{\pi ps}) = \sum_{k=1}^{n}(\frac{1-\pi_k}{\pi_k^2})y_k^2 + \sum_{k=1}^{n}\sum_{i\neq k}^{n}(\frac{\pi_{ki}-\pi_k\pi_i}{\pi_k\pi_i})\frac{y_k y_i}{\pi_{ki}},$$

where $\pi_k = np_k$ and π_{ki} is the joint inclusion probability of the unit k and i.

■

- *Stratified sampling with/without replacement in unequal probability selection*

  If quantitative and qualitative information are provided within a sampling frame, stratified random sampling with unequal probability selection may improve the efficiency of the estimator compared with SRS for a given sample size.

Theorem 2.9 *Stratified Unequal Probability Sampling with Replacement*

In stratified sampling with replacement and unequal probability selection, an unbiased estimator of τ is

$$\hat{\tau}_{pps}^{st} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{k=1}^{n_h} \frac{y_{hk}}{p_{hk}},$$

with a variance of

$$V_{st}(\hat{\tau}_{pps}) = \sum_{h=1}^{H} \frac{1}{n_h}[\sum_{k=1}^{N_h} \frac{y_{hk}^2}{p_{hk}} - \tau_h^2].$$

An unbiased estimator for the variance of the sample total is

$$v_{st}(\hat{\tau}_{pps}) = \sum_{h=1}^{H} \frac{1}{n_h(n_h-1)}[\sum_{k=1}^{n_h} \frac{y_{hk}^2}{p_{hk}^2} - n_h\hat{\tau}_{h,pps}^2],$$

where $\hat{\tau}_{h,pps}$ is the estimated total in stratum h.

■

Theorem 2.10 *Stratified Unequal Probability Sampling without Replacement*

In stratified sampling without replacement and unequal probability selection, an unbiased estimator of τ is

$$\hat{\tau}_{\pi ps}^{st} = \sum_{h=1}^{H}\sum_{k=1}^{n_h} \frac{y_{hk}}{\pi_{hk}},$$

with a variance of

$$V_{st}(\hat{Y}_{\pi ps}) = \sum_{h=1}^{H}\sum_{k=1}^{N_h}(\frac{1-\pi_{hk}}{\pi_{hk}})y_{hk}^2 + \sum_{h=1}^{H}\sum_{k=1}^{N_h}\sum_{i\neq k}^{N_h}(\frac{(\pi_{hki}-\pi_{hk}\pi_{hi})}{\pi_{hk}\pi_{hi}})y_{hk}y_{hi}.$$

An unbiased estimator for the variance of the sample total is

$$v_{st}(\hat{\tau}_{\pi ps}) = \sum_{h=1}^{H}\sum_{k=1}^{n_h}(\frac{1-\pi_{hk}}{\pi_{hk}^2})y_{hk}^2 + \sum_{h=1}^{H}\sum_{k=1}^{n_h}\sum_{i\neq k}^{n_h}(\frac{(\pi_{hki}-\pi_{hk}\pi_{hi})}{\pi_{hk}\pi_{hi}})\frac{y_{hk}y_{hi}}{\pi_{hki}},$$

where $\pi_{hk} = \frac{n_h x_{hk}}{\sum_{i=1}^{N_h} x_{hi}}$ and π_{hki} is the joint inclusion probability of the unit k and i in the stratum h.

■

d) *an approximation for unequal probability selection in without replacement scheme*

Since the joint inclusion probability in theorem 2.8 and 2.10 is complicated and difficult to compute, there are many strategies to ease computation. In this thesis, the *ordered estimates* as suggested by *Raj* (1956) is used.

Raj uses *Yates & Grundy draw by draw* procedure to select a sample to get ordered estimates as in the following procedure: (i) select the first unit in the sample with probability proportional to size z_k; (ii) select the second unit, without replacement, again with probability proportional to size; and (iii) draw until the sample size equals n. Ordered estimates are given in theorem 2.11 by *Raj*.

For more general discussion of *Yates & Grundy draw by draw* procedure and *Raj's ordered estimates* see *Brewer & Hanif* (1983) and *Govindarajulu* (1999).

Theorem 2.11 *Raj's approximation for unequal probability selection in without replacement sampling design*

In random sampling without replacement and unequal probability selection, an unbiased estimator of μ is

$$\hat{\mu}_D = \frac{1}{n}\sum_{k=1}^{n} t_k = \bar{t},$$

where $t_k = \frac{1}{N}\{y_1 + ... + y_{k-1} + \frac{y_k}{z_k}(1 - z_1 - z_2 - ... - z_{k-1})\}$ and z_k is an initial probability in the order of the k^{th} selection, with an unbiased variance estimate of

$$v(\hat{\mu}_D) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(t_k - \bar{t})^2.$$

■

In summary, estimation methods are used to draw conclusions about the population based on the information which has been gathered from the sample. There is

a direct association between probability sampling methods (2.3.1.6) and estimation procedures (2.3.2) in that the sampling design (2.3.3) itself determines the weights or expansion factors which are used to produce the estimates. Thus, the way to prove the properties of estimators such as unbiasedness depends on an inference approach about estimator.

2.4 Inference in Sample Surveys

In statistical inference a sample is used to draw inference about some aspect of the population from which the data are taken. Often the inference concerns the value of one or more unknown parameters, which describes some attribute of the population such as mean and its variance. This inference involves essentially three steps:

- Choice of a sampling plan.
- Choice of an estimator.
- Choice of a variance estimator and hence a confidence interval.

For example, if $SRSWOR$ is planned to conduct a survey, a simple estimator or a composite estimator is chosen to estimate a population mean. The variance of the estimator of the mean is automatically given by this sampling plan. However, if unequal probability sampling is used in $SRSWOR$, one of several approximations for the variance of the estimator of the mean must be chosen. There are two different types of inference based on different philosophies: (i) Frequentist approach and (ii) Bayesian approach.

i) *Frequentist* approach: The frequentist, classical or sampling theory approach is the most widely used approach. The theory makes the assumption that repeated samples of data from the population can be randomly taken under the same conditions as for our observed samples. Properties of point estimators and interval

estimators are all derived under this repeated sampling assumption. Frequentist inference works reasonably well in many circumstances, but in complicated situations it can break down and produce unreasonable results (*Garthwaite et al*, 1995).

In frequentist approach, there are three general inference approaches: (i.1) design-based approach also called randomisation approach, (i.2) model-based approach also called prediction approach and (i.3) model-assisted approach.

i.1) The *design-based* approach: Inference is based on the actual not the modelled or assumed sampling distribution. Repeated samples $\mathbf{S}$ are generated by the sampling design $p(\mathbf{S})$ with the values $y_1, ..., y_N$ held fixed. An essential property of this approach is that the sampling design determines how sampling variability is estimated. For example, any of the complexity due to the sampling scheme, such as multi-stage sampling, double sampling and so on, can be properly accounted for in estimation. This approach also is a nonparametric approach to inference since no assumptions about the distribution of random variables are assumed in making this inference. This is discussed more fully in section 2.4.1.

i.2) The *model-based* approach: Inference is based on the sampling distribution of estimator over repeated realisations $y_1, ..., y_N$ generated by the model, ξ, of the distribution of the $Y_1, ..., Y_N$ with the selected sample $\mathbf{S}$ held fixed. An essential property of this approach is the model determines how variability is estimated and the sampling design is irrelevant as long as the model holds. The accuracy of the inference depends on the validity of the model. If the model is wrong, the model-based estimate of the variance will underestimate the MSE since the model-based estimator is a biased estimator. In addition, there is no concept equivalent to complex sampling designs for the model-based approach and this approach is generally not appropriate for estimation. More details are in *Sarndal* (1978) and discussed further in section 2.4.2.

In this thesis, this approach cannot be used for proofs under nonresponse problem because it is more difficult to estimate the variance as well as the reason mentioned above. An alternative approach can then be used which is a combination between the assumed model step in the model-based approach and the inference step in the design-based approach. This leads to the model-assisted design-based approach or simply called model-assisted approach.

i.3) The *model-assisted* approach: To construct an estimator with good design-based properties, consistent with a plausible model for the variable of interest, Y, a model-assisted approach can be used. In this approach the finite population is assumed to be a realisation from some hypothetical superpopulation. Auxiliary information, X, can then be used by postulating models for the estimation of parameters of the finite population under consideration. The sampling weights are used to estimate the parameters and the sample design are used to estimate variances of the estimate. This inference is called a model-assisted approach since the model is used to specify the parameters of interest as in (i.2) and all inference is based on the survey design as in (i.1). More details are in section 2.4.3.

ii) *Bayesian* approach: In Bayesian inference, the parameter of interest, θ, is treated as a random variable rather than a constant as is generally the case in classical inference. This inference is based on *Bayes' Theorem* and consists of the following principal steps:

ii.1) Obtain the likelihood, $f(\mathbf{Y}|\theta)$, describing the process giving rise to the data $\mathbf{Y}$ in terms of the unknown parameters θ.

ii.2) Obtain the prior distribution, $f(\theta)$, expressing what is known about θ, prior to observing the data.

ii.3) Apply Bayes' theorem to derive the posterior distribution $f(\theta|\mathbf{Y})$ expressing what is known about θ after observing the data.

ii.4) Derive appropriate inference statements from the posterior distribution. These may include specific inferences such as point estimates, interval estimates or probabilities of hypotheses. If interest centres on particular components of θ, its posterior distributions is formed by integrating out the other parameters.

This form of inference differs from the classical form of frequentist inference in several respects, particularly the use of a prior distribution which is absent from classical inference. It represents the investigator's knowledge about the parameters before seeing the data. Classical statistics use only the likelihood. Consequently, in the Bayesian inference every problem is unique and is characterised by the investigator's beliefs about the parameters expressed in the prior distribution for the specific investigation. Thus, Bayesian inference may remove many of the mathematical and logical problems suffered from the frequentist approach. However, it suffers from practical difficulties, in particular how to choose the prior distribution of the parameters. Moreover, there is no need to invoke the possibility of repeating sampling. More details are in sections 2.4.4.

2.4.1 Design-Based Approach with Complete Data

Let $\mathbf{Y}$ be the variable of interest and $\mathbf{I}$ be the indicator as defined in section 2.3.1.1. In design-based approach, sample units are selected by probability sampling which is characterised by the following two properties:

- The sampling distribution, denoted by $f(\mathbf{I}|\mathbf{Y})$ (see below for example), is determined by the sampling designer before any y values are known.
- Every unit has a positive known probability of selection, i.e.

$$\pi_k = P(I_k = 1|\mathbf{Y}) > 0 \text{ for all } k = 1, ..., N$$

For example, in $SRSWOR$ with sample size n, the sample selection process is characterised by the conditional distribution of $\mathbf{I}$ given $\mathbf{Y}$,

$$f(\mathbf{I}|\mathbf{Y}) = f(\mathbf{I}) = \begin{cases} 1/\binom{N}{n} & \text{if } \sum_{k=1}^{N} I_k = n \\ 0 & \text{otherwise,} \end{cases}$$

so that every possible of n distinct units in the population has the same probability of being selected as the sample, $1/\binom{N}{n}$. As a consequence of these properties, the probability that any given unit appears in the sample, π_k, is n/N. Thus, for example, as stated in theorem 2.2, $\hat{\mu}_{srs} = \frac{1}{n}\sum_{k=1}^{n} y_k$, can be proven to be an unbiased estimator under design-based approach as follow:

$$\begin{aligned} E(\hat{\mu}_{srs}) = E(\tfrac{1}{n}\textstyle\sum_{k=1}^{n} y_k) &= E(\frac{1}{n}\sum_{k=1}^{N} I_k y_k) \\ &= \frac{1}{n}\sum_{k=1}^{N} y_k E(I_k) \\ &= \frac{1}{n}\frac{n}{N}\sum_{k=1}^{N} y_k \\ &= \mu. \end{aligned}$$

In addition when a qualitative auxiliary information, X, is available in the sampling frame and $STWOR$ is adopted for use, sample selection process is then characterised by

$$f(\mathbf{I}|\mathbf{Y}, \mathbf{X}) = f(\mathbf{I}, \mathbf{X}) = \begin{cases} 1/\binom{N_h}{n_h} & \text{if } X \in S_h \text{ and } \sum_{k=1}^{N} I_k = n_h \\ 0 & \text{otherwise.} \end{cases}$$

In this case an unbiased estimator of the mean estimate is

$$\hat{\mu}_{st} = \textstyle\sum_{h=1}^{H} \frac{N_h}{N}\hat{\mu}_{srs},$$

since sampling is done independently in each strata.

2.4.2 Model-Based Approach with Complete Data

Let $y_1, ..., y_N$ be realisations of the random variables $Y_1, ..., Y_N$ generated from a model, ξ, of the distribution of the relevant $Y_1, ..., Y_N$. This joint probability distribution of $Y_1, ..., Y_N$ supplies the link between units in the sample and units not in the sample in the model-based approach. The samples y_k, $k \in S$ are used to predict the unobserved values $y_k \notin S$. Problems in finite population sampling may be thought of as a prediction problem. If a different model is chosen to explain the variable of interest, Y, the variance of the estimator may differ because it depends on the model used.

For example, in $SRSWOR$ with a sample size n, an unbiased total estimator and its variance, $\hat{\tau}_{srs} = \frac{N}{n}\sum_{k\in S} y_k$ and $V(\hat{\tau}_{srs})$, are different under the model-based approach, with two different assumed models for the distribution of $Y_1, ..., Y_N$. This point is illustrated as follows:

i) Assume that the model $Y_1, ..., Y_N$ are independent identically distributed with $E_\xi(Y_k) = \mu$ and $V_\xi(Y_k) = \sigma^2$, where μ and σ^2 are unknown population parameters. Under this model, the estimator of the population total is $\hat{\tau}_{srs} = \frac{N}{n}\sum_{k=1}^{n} y_k$ with a variance

$$V_\xi(\hat{\tau}_{srs}) = N^2(1 - \frac{n}{N})\frac{\sigma^2}{n}. \tag{2.1}$$

In practice, if the model described above were used, the population variance σ^2 can be estimated by the sample variance s^2. Thus, with this model the design-based approach and the model-based approach lead to the same variance estimate.

ii) Assume that the model Y is explained by a simple linear regression model, $Y_k = \beta_0 + \beta_1 X_k + \epsilon_k$, where β_0 and β_1 are unknown constant, X_k are known auxiliary information for each unit k, and ϵ_k are independent random variable

with zeros mean and variance σ^2. The estimator of the total estimate is $\hat{\tau}_{srs} = N(\hat{\beta}_0 + \hat{\beta}_1\mu_x)$ with a variance

$$V_\xi(\hat{\tau}_{srs,reg}) = N^2\sigma^2[\frac{1}{n} + \frac{(\mu_x - \hat{\mu}_x)^2}{\sum_{k=1}^n (x_k - \hat{\mu}_x)^2}]. \tag{2.2}$$

Thus, the total estimator, $\hat{\tau}_{srs}$, is N times the predicted value of Y at μ_x under the model. From the regression theory (*Sen & Srivestava*, 1990), the variance of $\hat{\beta}_0 + \hat{\beta}_1\mu_x$ is

$$\sigma^2[\tfrac{1}{n} + \tfrac{(\mu_x - \hat{\mu}_x)^2}{\sum_{k=1}^n (x_k - \hat{\mu}_x)^2}].$$

Thus, the variance of the total estimate is N^2 times $V(\hat{\beta}_0 + \hat{\beta}_1\mu_x)$ as shown in 2.2 . Again, in practice a sample variance s^2 is used for σ^2.

2.4.3 Model-Assisted Approach with Complete Data

To illustrate the model-assisted approach, the regression model as in section 2.4.2 is considered. An auxiliary information, X may improve on the estimator of the total estimate Y, $\hat{\tau}_y = \frac{N}{n}\sum_{k=1}^n y_k$, through the regression model. Assume that the true population total τ_x is known and thus can be used to adjust the estimate $\hat{\tau}_y$ as

$$\hat{\tau}_{ygreg} = \hat{\tau}_y + (\tau_x - \hat{\tau}_x)\hat{\beta}_1,$$

which is called the *generalised regression estimator* of the population total, where $\hat{\beta}_1 = \sum_{k=1}^n (y_k - \hat{\mu}_y)(x_k - \hat{\mu}_x)/\sum_{k=1}^n (x_k - \hat{\mu}_x)^2$ and $\hat{\tau}_x = \frac{N}{n}\sum_{k=1}^n x_k$.

However, the variance of the estimator of the total is computed by using design-based inference as (*Cochran*, 1977):

$$\begin{aligned} V(\hat{\tau}_{ygreg}) &= V[\hat{\tau}_y + (\tau_x - \hat{\tau}_x)\hat{\beta}_1] \\ &\approx N^2(1 - \frac{n}{N})\frac{\sigma_y^2}{n}(1 - \rho^2) \\ &= N^2(1 - \frac{n}{N})\frac{\sigma_e^2}{n}, \end{aligned}$$

where a residual unit k, $e_k = y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k$, $\hat{\beta}_0 = \hat{\mu}_y - \hat{\beta}_1 \hat{\mu}_x$ and ρ is the population correlation between Y and X.

If the model is a good one, the variability in the residuals is expected to be smaller than the variability in the original observations. The generalised regression estimator, $\hat{\tau}_{ygreg}$, will be more efficient than the simple estimator $\hat{\tau}_y$. For example, in an SRS,

$$\hat{V}_{srs}(\hat{\tau}_y) = \frac{N^2}{n}(1 - \frac{n}{N}) \sum_{k=1}^{n} \frac{(y_k - \hat{\mu}_y)^2}{n-1},$$

but

$$\hat{V}_{srs}(\hat{\tau}_{ygreg}) = \frac{N^2}{n}(1 - \frac{n}{N}) \sum_{k=1}^{n} \frac{e_k^2}{n-1}.$$

If the residuals tend to be smaller than the deviation of y_k about the sample mean, then the estimated variance is smaller for the generalised regression estimator. Note however that generalised regression (GREG) estimation may lead to negative weights.

2.4.4 Bayesian Approach with Complete Data

The Bayesian modelling approach to sampling with complete data is to treat $\mathbf{I}$ and $\mathbf{Y}$ in the population as realisations of random variables with joint distribution $f(\mathbf{Y}, \mathbf{I}, \theta | \mathbf{X})$. Here, as before, $\mathbf{I}, \mathbf{X}$ and $\mathbf{Y}$ are the sample indicator, auxiliary information and study variable respectively and θ is a vector of parameters of interest.

A finite population $Y_1, ..., Y_N$ is considered as a random sample of N independent observations on a random variable $\mathbf{Y}$ whose distribution is $f(\mathbf{Y}|\mathbf{X}, \theta)$. Write $Y^T = (\mathbf{Y_{inc}}, \mathbf{Y_{exc}})^T$, where $\mathbf{Y_{inc}}$ is the set of Y values included in the sample, and $\mathbf{Y_{exc}}$ is the set of Y values excluded from the sample. The data observed in the absence of nonresponse are then $\mathbf{Y_{inc}}, \mathbf{I}$ and $\mathbf{X}$.

Inference for a population quantity, e.g., the population mean $\mu_y = \frac{1}{N}\{n\hat{\mu}_{inc} + (N-n)\hat{\mu}_{exc}\}$ where $\hat{\mu}_{inc}$ and $\hat{\mu}_{exc}$ are the sample mean and non-sample mean respectively, is obtained by:

i) Finding the marginal posterior distribution for the unobserved finite population units $\mathbf{Y_{exc}}$ given the sample $\mathbf{Y_{inc}}, \mathbf{I}$ and $\mathbf{X}$:

$$f(\mathbf{Y_{exc}}|\mathbf{Y_{inc}}, \mathbf{I}, X) = f(\mathbf{Y_{exc}}|\mathbf{Y_{inc}}, \mathbf{X}), \tag{2.3}$$

where the sampling mechanism is assumed ignorable. This is more for convenience and ease in deriving results. More details are in *Little* (1982) and *Rubin* (1976, 1987).

ii) Under the ignorable sampling mechanism mentioned above in (i), finding the conditional posterior distribution expectation and variance for μ_y given $\mathbf{Y_{inc}}$ and $\mathbf{X}$ by using the marginal predictive distribution from equation 2.3 and the posterior of θ given $\mathbf{Y_{inc}}$ and $\mathbf{X}$, which is assumed known, $f(\theta|\mathbf{Y_{inc}}, \mathbf{X})$:

$$E(\mu_y|\mathbf{Y_{inc}}, \mathbf{X}) = \tfrac{1}{N}\{n\hat{\mu}_{inc} + (N-n)E(\hat{\mu}_{exc}|\mathbf{Y_{inc}}, \mathbf{X})\},$$

where

$$\begin{aligned} E(\hat{\mu}_{exc}|\mathbf{Y_{inc}}, \mathbf{X}) &= \frac{1}{N-n}\sum_{k\notin S} E(Y_k|\mathbf{Y_{inc}}, \mathbf{X}), \\ S &= \{k : I_k = 1\}, \\ E(Y_k|\mathbf{Y_{inc}}, X) &= \int\cdots\int Y_k f(\mathbf{Y_{exc}}|\mathbf{Y_{inc}}, X)d\mathbf{Y_{exc}}, \end{aligned}$$

and

$$V(\mu_y|\mathbf{Y_{inc}}, \mathbf{X}) = (\tfrac{N-n}{N})^2 V(\hat{\mu}_{exc}|\mathbf{Y_{inc}}, \mathbf{X}).$$

For example, in $SRSWOR$ with sample size n, let Y be the study variable measured for units in the sample. If the ignorable mechanism is assumed and the

superpopulation is assumed normal with μ and variance σ^2 and 'diffuse' priors on μ and σ^2 are used, it is shown by *Palit & Gutman* (1972) that the posterior expectation of the finite population mean is the sample mean,

$$E(\mu_y|\mathbf{Y_{inc}}) = \hat{\mu}_{inc},$$

and that the posterior variance of the finite population mean is

$$V(\mu_y|\mathbf{Y_{inc}}) = \frac{N-n}{N}\frac{\sigma^2}{n},$$

if σ^2 is known, and

$$V(\mu_y|\mathbf{Y_{inc}}) = \frac{N-n}{N}\frac{u^2}{n},$$

where $u^2 = \frac{1}{n-3}\sum_{k=1}^{n}(y_k - \hat{\mu}_{inc})^2$ when σ^2 is unknown.

In this thesis the ignorable sampling mechanism is assumed. Furthermore, it should be noted that even with ignorable sampling mechanisms the sampling design does affect the impact of specification errors and thus influence the choice of model indirectly but these aspects are not considered here. For more details see *Rubin* (1987). However, not all sampling mechanisms may be ignorable. In these cases inference based on equation 2.3 may be subject to bias and also the full marginal predictive model is generally hard to specify unless exclusion from the sampling is determined by a known mechanism, such as censoring with known censoring points.

2.5 Sampling and Non-sampling Errors

The typical survey objective is to estimate a descriptive population quantity called a parameter. This leads to consideration of how inference can be made about the population using information contained in a sample. In this step, the probability distribution of the sample, called the *sampling distribution*, is used. Knowledge of

this distribution allows choice of proper inference-making procedures and to attach measures of goodness to such inference. In order to get a valid inference, ideally:

i) The population from which the sample is selected is the population of interest, and that all selected units can be measured.

ii) These measurements give the true value of any variable or category of any attribute of interest.

However, these assumptions are frequently violated in practice. This leads to errors. The amount by which the estimate differs from the true value of the population parameter is called *total survey errors*. These errors can be divided into two basic types of errors which arise: (i) sampling errors and (ii) non-sampling errors.

i) *Sampling errors* (SE) are the errors attributed to studying only a subset of the represented population. These errors, at least in the case of variance or standard error, can be measured theoretically in probability sampling. They can be estimated from the sample data when probability sampling is used and in general this tends to decrease as $1/\sqrt{n}$.

In principle, sampling errors can be made small by the choice of a sufficiently large and well-deployed sample, such as stratified random sample or unequal probability sample.

ii) *Non-sampling errors* (NSE) are any errors that cannot be attributed to the sample-to-sample variability. These errors are present both in sample surveys and in censuses and can occur at every stage of planning and execution of the census or survey.

Non-sampling errors may be broadly classified into three areas that correspond roughly to the types of activities in a survey (*Lessler & Kalsbeek*, 1992): (i) constructing a sampling frame; (ii) locating sample members and soliciting their participation in the survey; and (iii) collecting data and converting it into machine-readable

form. The survey errors associated with these three basic activities are called frame errors, nonresponse errors and measurement errors respectively. Another classification for non-sampling errors are to divide them into three categories corresponding to the three stages of census or survey work (*Murthy*, 1967): (i) planning stage leading to specification errors, (ii) field work stage leading to ascertainment errors and (iii) tabulation stage leading to tabulation errors. *Thompson* (1997), on the other hand, divides non-sampling errors into coverage errors, nonresponse errors, and response or measurement errors, and write *NSE* as *coverage errors+nonresponse errors+measurement errors.*

In general, if a randomisation procedure is used in selecting the sample, the extent of sampling errors is much easier to estimate than the extent of non-sampling errors. Moreover, non-sampling errors are often left underestimated or unacknowledged in reports of surveys because it is very hard to examine errors in every step of the survey. The only way to control non-sampling errors is to exercise great care in the planning and execution stage of the survey.

If non-sampling errors can be assumed to be *random* with zero mean, they do not cause bias or other complications in the estimation. These type of non-sampling errors are called *ignorable non-sampling errors.* When non-sampling errors are not random with zero mean, they can be harmful as they are likely to cause bias in the estimation and they are called *nonignorable non-sampling errors.* More details with non-sampling errors are in *Biemer* (1991) and *Groves* (1984).

In the practice non-sampling errors may occurs with nonresponse errors, frame errors and measurement errors but the theory in this thesis, nonresponse errors are assumed to be the only non-sampling errors and this study focuses on ways to reduce the effect of nonresponse.

2.6 Nonresponse

Once the sample is selected, field work begins and an attempt is made to collect the desired data from all enumeration units selected in the sample. Unfortunately, it is rarely possible to achieve the complete data set from all units sampled. For some units, the sample survey may have obtained no information at all, and for other units, the survey may have obtained information on some, but not all of the items in the units. The former type of nonresponse is called *unit nonresponse*, while the latter is called *item nonresponse*. Unit nonresponse arises because the unit is unwilling to answer, unable to participate, not available, or not traceable. Item nonresponse arises because one or more items in the unit are not answered, without a clear answer, or not acceptably answered (*Kviz*, 1998).

Both unit nonresponse and item nonresponse are a major cause of inaccuracy in sample surveys. Both types of nonresponse are very difficult to avoid in sampling. The increase over the years in the use of sample surveys to provide information for purposes of decision making, and the increasing difficulty of obtaining high response rates in sample surveys, has resulted in considerable attention being paid to nonresponse problems. Thus a wide variety of techniques for dealing with nonresponse in sample surveys have been developed. The impact of nonresponse on the bias of estimates obtained from sample surveys is discussed below in section 2.6.1. Some methods that have been used to reduce errors due to nonresponse are discussed in section 2.6.2. Mechanisms of nonresponse are described in section 2.6.3. Finally, section 2.6.4 presents on inference approach to nonresponse.

2.6.1 Bias due to Nonresponse

The main problem caused by nonresponse is potential bias of population estimates. Suppose a population is divided into two artificial "strata" of respondents and non-

respondents. Let N_R be population respondent units that would respond if they were chosen to be in the sample and let μ_R be the responding population mean. Similarly, the N_{NR} population nonrespondents are the units that would not respond and μ_{NR} is the nonrespondent population mean. The following notation is used to prove the bias.

Let $W_R = N_R/N$ and $W_{NR} = N_{NR}/N$, so that W_{NR} is the proportion of nonresponse in the population. When the field work is completed, data from respondent phase are collected but no data from the nonrespondent phase. *Cochran* (1977) shows the amount of bias in the sample mean is:

$$E(\hat{\mu}_R) - \mu = \mu_R - \mu = \mu_R - (W_R\mu_R + W_{NR}\mu_{NR}) = W_{NR}(\mu_R - \mu_{NR}).$$

Thus, the amount of bias is the product of the proportion of nonresponse and the difference between the mean in the two phases. Since the sample provides no information about μ_{NR}, the size of the bias is unknown unless bounds can be placed on μ_{NR} from some sources other than the sample data. The bias cannot be measured exactly. Even though it is hard to get objective measure of the bias, it is relatively simple to quantify the extent of the nonresponse. Different measures of the nonresponse are usually found in the quality declarations that statistical agencies and survey institutes often publish together with the results. For example, the guideline for reporting response rates in *Hidiroglou & Drew* (1993) and *Lessler & Kalsbeck* (1992) provide a sensible solution for reporting response rates and relative measures of nonresponse such as nonresponse rate, completion rate, refusal rate, etc. The end user should take this information into account when judging the credibility of the results.

The bias due to nonresponse tends to small if either (i) the mean for the nonrespondents is close to the mean for the respondents or (ii) there is little nonresponse (W_{NR} is small). However, it is no assurance of (i) since there is no data for the

nonresponse. Minimising the nonresponse rate should be the first point of effort in controlling nonresponse. More details for minimising nonrespondents are in *Rossi et al* (1983), *Rubin et al* (1995) *Hidiroglou et al* (1993) and *Sudman* (1998).

2.6.2 Dealing with Nonresponse

There are generally three ways of dealing with unit nonresponse: (i) by planning of the survey before data collection, (ii) by using special techniques during data collection and (iii) by making model assumptions about the response mechanism after data collection.

In this thesis, ways of dealing with nonresponse by (ii), special techniques during data collection and (iii) special techniques with response model assumption after data collection are considered. These methods are discussed in depth in chapter 3, 4 and 5. For completeness this section briefly describes (i) planning of the survey that can be taken before data collection and the special efforts during data collection.

2.6.2.1 Planning of the Survey

The ideal survey has no nonresponse. To come close to this ideal requires careful planning and often considerable expense. The nonresponse is affected by a number of the operations that define the survey. Special effort must be made at the planning stage to foresee how alternative survey operations may influence the response.

The selection, training and supervision of interviewers are factors of great importance. The choice of data collection method such as personal interview, telephone interview, mail inquiry, and so on, is important, as is the length and content of the questionnaire or schedule. In a repeated survey, the frequency with which respondents are asked to participate must be considered. Response rates are often adversely affected by a heavy response burden.

Effective measures should be taken to reduce the nonresponse to insignificant levels so that any remaining nonresponse cause little or no harm to the validity of

the inference.

2.6.2.2 Special Efforts

In surveys with personal interviews, the first contact with a potential respondent may be unsuccessful for a variety of reasons. For example, no one may be available, the person selected may be sick, or the interview may be broken off before completion. If the first attempted contact results in too many unsuccessful interviews, it is common to make one or more *callbacks* at more convenient times, perhaps using interviewers with special skills and experience. The nonresponse can thereby be considerably reduced if not eliminated. Callbacks can also give valuable information about the selective effects of nonresponse. For example, *Rao* (1983) shows that adult members of households with young children are more likely to be at home when the interviewer calls, so that if efforts to elicit a response were stopped after a single call, the response set would tend to over represent households with small children. Thus if the target population is all households, the estimates for variables correlated with number of children will tend to be biased.

In mail surveys, *follow-up* letters in combination with telephone interviews and personal interviews are often used as a callback technique. Often two or three reminders are mailed to obtain as many responses as possible from the mail phase. As with personal interviews, the propensity to answer at different stages of callback is often correlated with the study variable. *Rao* (1983) gives as an example the 1950 survey of North Carolina Fruit Growers where late respondents and nonrespondents differed considerably from those responding at the first mailing with a strong association between propensity to respond early and large size of farm.

In practice, callbacks or follow-ups have to stop after a few attempts. Furthermore, the reduction in the mean square error of the survey estimates will often be small compared to the cost of further callbacks. To counter this, *Deming* (1953)

and *Thomsen & Siring* (1983) developed a model for the study for different number of callback strategies. As an alternative, *Rao* (1983) also suggests the methods, proposed by *Politz & Simmons* (1949), which a procedure is based on weighting the responses with estimated response probabilities to deal with the problem "not-at-home". For some surveys, this is more cost efficient than repeated callbacks. More details with callbacks and follow-ups are in *Dillman* (1998).

2.6.3 Mechanisms of Nonresponse

Most surveys have some residual nonresponse even after careful design and follow-up of nonresponse. Nearly all methods for reducing the impact of nonresponse are model-based. Population members can be divided into two fixed strata of would-be respondents and would-be nonrespondents. To adjust for nonresponse that remains after all other measures have been taken, the response or nonresponse of unit k can be assumed to be a random variable. Let the random variable R equal one if unit k responds, otherwise zero. The probability that a unit selected for the sample will respond, $\phi_k = P(R_k = 1)$, is unknown but assumed non-zero. After sampling, the values of R_k are known for units selected in the sample. A value for Y_k is recorded if r_k, the realisation of R_k, is 1.

There are generally three types of missing data proposed by *Little & Rubin* (1987) as: (i) missing completely at random, (ii) missing at random given covariates and (iii) nonignorable nonresponse. More details are in *little* (1998) and *Rubin* (1977, 1983).

i) *Missing Completely at Random*: The missing data are called missing completely at random (*MCAR*) if the response probability, ϕ_k, does not depend on x_k, y_k, or the survey design. Then, the respondents are representative of the selected sample and the nonrespondents are assumed to be selected at random

from the sample. For example, if a simple random sample of size n were taken and the response probabilities ϕ_k are all equal with the events, $\{R_k = 1\}$, being conditionally independent of each other and of the sample selection process, then the data are *MCAR*. In this case the sample mean of respondents, $\hat{\mu}_R$, is approximately an unbiased estimator of the population mean. In general when ϕ_k are equal the *MCAR* mechanism is implicitly adopted and the nonrespondents are ignored.

MCAR is discussed with nonresponse for some sampling designs in chapter 3, 4 and 5 and with a naive model in chapter 4 and 5.

ii) *Missing at Random Given Covariates* or *Ignorable Nonresponse*: If ϕ_k does not depend on y_k but is determined by x_k, the data are called missing at random (*MAR*) or sometimes ignorable nonresponse. Since the values of X_k are known for all sample units, a model can be used to estimate y_k for the nonresponse units. Thus the nonresponse can be ignored since the model compensates for it.

MAR is discussed with a *RHG* model in chapter 4 and 5.

iii) *Nonignorable Nonresponse*: If the probabilities of nonresponse depends on the value of a response variable and cannot be completely explained by values of $X's$, the nonresponse is called *nonignorable*. Models may help with the nonresponse situation since the nonresponse probability may be approximated by using x_k. However, the adjustment for the nonresponse will be approximate.

Nonignorable nonresponse is discussed with multiple imputations in chapter 5.

Lohr (1999) summarises how to distinguish the nonresponse mechanism as:

The way to think about the type of nonresponse is that the probabilities of

responding, θ_i, are useful for checking about nonresponse mechanism. Unfortunately, they are unknown, so we do not know for sure which type of nonresponse is present. $MCAR$ and MAR can be sometimes distinguished between them by fitting a model that attempts to predict the observed probabilities of response for subgroups from known covariates. If the coefficients in a logistic regression model are significantly different from zero, the missing data are likely not $MCAR$. Distinguishing between MAR and nonignorable nonresponse is more difficult.

2.6.4 Inference Approach with Nonresponse

Inference discussed in section 2.4 is concerned with a complete sample. In this section inference is based on nonresponse as well as information from the responses in the sample. When a sample is not complete, the inference described in section 2.4 cannot be used properly. There are generally three ways of dealing with nonresponse in inference: (i) Quasi-randomisation approach, (ii) Model-Assisted approach and (iii) Bayesian approach. These approaches are studied in this thesis.

2.6.4.1 Quasi-Randomisation Approach with Nonresponse

The key ingredient of the randomisation approach in section 2.4.1 is the known probability distribution, $f(\mathbf{I}|\mathbf{Y},\mathbf{X})$ governing which sample units are selected. However, when some of the data are missing, using this distribution which assumes full response will lead to a biased estimate of a parameter. For example, suppose that n sample units are selected by SRS and let I_k and R_k be the sampling and response indicators as defined in section 2.3.1.1 and 2.6.3 respectively. Then, the value of Y_k is recorded if and only if $R_k = I_k = 1$. Section 2.6.1 shows that the estimator based on randomisation inference approach with nonresponse for the mean estimate is biased. Thus, it is difficult to define an unbiased estimator which is a function

of the recorded values of $\mathbf{Y}$ with respect to the distribution of $\mathbf{I}$ and these does not appear to be in any part of sampling textbooks or journals. An approximation was given by *Oh & Scheuren* (1983) is used here. They proposed a quasi-randomisation approach which formulates a distribution for $\mathbf{R}$. Thus, the quasi-randomisation assumes: (i) a known distribution $f(\mathbf{I}|\mathbf{Y},\mathbf{X})$ of a sample selection as for complete survey data as before and (ii) an assumed distribution for the response indicator $\mathbf{R}$ given $\mathbf{I}$, $\mathbf{Y}$ and $\mathbf{X}$. More details are fully discussed in *Little & Rubin* (1987).

However, the strong assumption underlying $f(\mathbf{R}|\mathbf{I},\mathbf{Y},\mathbf{X})$ is that $\mathbf{R}$ is independent of $\mathbf{I}$, $\mathbf{Y}$ and $\mathbf{X}$. This assumption is often unrealistic in practice. The quasi-randomisation inference is used for weighting adjustment procedures in chapter 4 and random imputation in chapter 5.

2.6.4.2 Model-Assisted Approach with Nonresponse

When sample data is missing, the problem of estimating the population total will differ from predictive inference with complete data. Since the total population can be expressed as

$$\tau = \textstyle\sum_{k\in S_1} y_k + \sum_{k\in S_0} y_k + \sum_{k\in \tilde{S}} y_k,$$

where S_1 denotes the m response sampled unit labels for which y is available; S_0 denotes the $n-m$ nonresponse sampled unit labels for which y is not available; and $\tilde{S}$ denotes the $N-n$ non-sampled unit labels. The problem of estimating the total is that of estimating the sum of the sample values not observed and the sum of the non-sampled values. However, this problem can be reduced into one of estimating the sum of the sample values not observed in the sample as:

$$\hat{\tau} \quad = \quad \frac{N}{n}\Big(\sum_{k\in S_1} y_k + \sum_{k\in S_0} \hat{y}_k\Big).$$

In the model-assisted approach, a regression model is one of methods that is used to impute the missing data, e.g., a simple case is considered as follow and in chapter 5.4:

$$\hat{\tau} = \frac{N}{n}[\sum_{k=1}^{m} y_k + \sum_{k=m+1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_k)],$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are sample regression coefficients based on the m responded samples.

$\hat{\tau}$ is an unbiased estimator of the population total τ only if the regression model is true. For variance estimation, a randomisation approach is used to infer the variance of the estimator of the total. More details of a linear regression model used for dealing with nonresponse are discussed in section 5.2.3.

2.6.4.3 Bayesian Approach with Nonresponse

The Bayesian modelling approach to sampling with nonresponse is to treat $\mathbf{I}$, $\mathbf{R}$, $\mathbf{Y}$ and θ in the population as realisations of random variables with joint distribution $f(\mathbf{Y}, \mathbf{I}, \mathbf{R}, \theta|\mathbf{X})$. Here, as before, $\mathbf{I}, \mathbf{R}, \mathbf{X}$ and $\mathbf{Y}$ are the sample indicator, response indicator, auxiliary information and variable of study respectively and θ is a vector of parameter of interest.

A finite population $Y_1, ..., Y_N$ is considered as a random sample of N independent observations on a random variable $\mathbf{Y}$ whose distribution is $f(\mathbf{Y}|\mathbf{X}, \theta)$. Inference for a population quantity, e.g., the population mean $\mu_y = \frac{1}{N}\{m\hat{\mu}_{obs} + (N-m)\hat{\mu}_{nobs}\}$ where $\hat{\mu}_{obs}$ and $\hat{\mu}_{nobs}$ are the response sample mean and non-observed mean respectively, is obtained by:

i) Finding the marginal posterior distribution for the unobserved finite population units $\mathbf{Y_{nobs}}$ given the response sample $\mathbf{Y_{obs}}, \mathbf{I}, \mathbf{R}$ and $\mathbf{X}$:

$$f(\mathbf{Y_{nobs}}|\mathbf{Y_{obs}}, \mathbf{I}, \mathbf{R}, \mathbf{X}) = f(\mathbf{Y_{exc}}|\mathbf{Y_{inc}}, \mathbf{X}), \tag{2.4}$$

where the sampling mechanism is assumed ignorable. More details are in *Little* (1982) and *Rubin* (1976, 1987).

ii) Under the jointly ignorable nonresponse and sampling mechanisms (see section 2.6.3 and 2.4.1 respectively), finding the conditional posterior expectation and

variance for μ_y given $\mathbf{Y_{obs}}$ and $\mathbf{X}$ by using the marginal predictive distribution from equation 2.4 and the posterior distribution of θ given $\mathbf{Y_{obs}}$ and $\mathbf{X}$, which is assumed known, $f(\theta|\mathbf{Y_{obs}}, \mathbf{X})$:

$$E(\mu_y|\mathbf{Y_{obs}}, \mathbf{X}) = \tfrac{1}{N}\{m\hat{\mu}_{obs} + (N-m)E(\hat{\mu}_{nobs}|\mathbf{Y_{obs}}, \mathbf{X})\},$$

where

$$\begin{aligned} E(\hat{\mu}_{nobs}|\mathbf{Y_{obs}}, \mathbf{X}) &= \frac{1}{N-m}\sum_{k\notin S_1} E(Y_k|\mathbf{Y_{obs}}, \mathbf{X}), \\ E(Y_k|\mathbf{Y_{obs}}, \mathbf{X}) &= \int\cdots\int Y_k f(\mathbf{Y_{nobs}}|\mathbf{Y_{obs}}, \mathbf{X}) d\mathbf{Y_{nobs}}, \end{aligned}$$

and

$$V(\mu_y|\mathbf{Y_{obs}}, \mathbf{X}) = (\tfrac{N-m}{N})^2 V(\hat{\mu}_{nobs}|\mathbf{Y_{obs}}, \mathbf{X}).$$

More details are in *Little & Rubin* (1987).

Nearly all analytic procedures for handling nonresponse in sample survey practice effectively assume ignorable nonresponse because of its convenience. Even putting aside this aspect, many technical adjustments for nonresponse require ignorability of the nonresponse mechanism. $MCAR$ and MAR while mathematically tractable are difficult to check in practice. To check nonresponse mechanism logistic regression is rquired. See more details in *Lohr* (1999). However, not all nonresponse and sampling mechanisms are ignorable. In cases where they are not inference based on equation 2.4 may be subject to bias. The full marginal predictive model is also generally hard to specify. For more details see *Rubin* (1987). Bayesian inference approach is used with multiple imputation in chapter 5.

2.7 Comparing the Sampling Designs

When a sampling designer makes a decision about a sampling design, the search for the best design is made easier if, as has been done in some cases, the comparison

of sampling designs can be given a precise mathematical formulation. For example, what is the best sampling design in a specified class of designs if the objective is to minimise variance or to minimise variance for a given cost of sampling? The sampling designer has to choose not only the type of sampling techniques but also the estimators to find the best sampling design.

To compare different sampling designs in terms of efficiency, a measure of the efficiency of a sampling design is often obtained by using the ratio between the variance or mean square error of two survey designs and usually, simple random sampling with or without replacement is chosen as a reference. *Cornfield* (1951) suggested measuring the efficiency of a sampling design by using the ratio of the variance that would be obtained from simple random sampling design to the variance obtained from other designs with the same size. Two variations on this idea are commonly used (i) design effect and (ii) misspecification effect.

2.7.1 Design Effect

The *design effect* or *deff*, suggested by *Kish* (1965), is the reciprocal of *Cornfield's ratio*. It is used to summarise the effect of the sampling design on the variance of the estimate as:

$$deff(\hat{\theta}^*) = \frac{v_{true}(\hat{\theta}^*)}{v_{srs}(\hat{\theta})},$$

where $v_{srs}(\hat{\theta})$ is a sample variance of a simple estimator $\hat{\theta}$ (as defined in section 2.3.2.1) in SRS and $v_{true}(\hat{\theta}^*)$ is the sample variance of the estimator $\hat{\theta}^*$ used in the survey design.

For example, if the regression estimator of the total estimate is used in stratified random sampling, then design effect is defined as

$$deff(\hat{\tau}_{reg}) = \frac{v_{st}(\hat{\tau}_{reg})}{v_{srs}(N\hat{\mu})},$$

where $N\hat{\mu}$ and $\hat{\tau}_{reg}$ are the estimated population total on the simple estimate and the regression estimate respectively.

Thus, for example, a *deff* of less than one means a sampling design that is more efficient than SRS.

2.7.2 Misspecification Effect

The *misspecification effect* or *meff*, suggested by *Skinner et al* (1989), is an alternative method of comparing the efficiency of sampling designs. Let $v_0 = v_{IID}(\hat{\theta})$ be an estimator of the variance of $\hat{\theta}$ derived under the model assumption that observations are identical and independently distributed (IID), or equivalently under the design assumption that the sample is selected by SRS. The misspecification effect for the variance estimator of $\hat{\theta}^*$ is given by

$$meff(\hat{\theta}^*) = \frac{v_{true}(\hat{\theta}^*)}{E_{true}(v_o)}.$$

For example, if the regression estimator of the total estimate is used to compare the efficiency between stratified random sampling and simple random sampling, $\hat{\theta}^*$ and $\hat{\theta}$ are both $\hat{\tau}_{reg}$, then the estimated misspecification effect is

$$\widehat{meff}(\hat{\tau}_{reg}) = \frac{v_{ST}(\hat{\tau}_{reg})}{v_{SRS}(\hat{\tau}_{reg})}.$$

Thus, for example, a *meff* of less than one means a sampling design that is more efficient than SRS.

However, if v_o is assumed with design-based on SRS, then $E_{true}(v_o)$ equal $V_{SRS}(\hat{\theta})$ and $\hat{\theta}$ equals $N\hat{\mu}$. Then by the above example (*Skinner et al*, 1989),

$$\widehat{meff}(\hat{\tau}_{reg}) = \frac{v_{ST}(\hat{\tau}_{reg})}{v_{SRS}(N\hat{\mu})} = deff(\hat{\tau}_{reg}).$$

In general, however, *meff* does not equal to *deff*. Because of above result, *deff* is used in thesis to compare sampling designs.

Chapter 3

Nonrespondent Subsampling

Nonresponse is almost inevitable in most surveys. Callbacks or follow-ups are often used to compensate for this problem by eliminating or at least greatly reducing nonresponse. In theory, these techniques work well, but in practice there are still problems. A long series of callbacks and follow-ups may be costly and time consuming. Nonresponse may still be unacceptably high after the final call or follow-up letter especially in mail survey. Several authors have suggested methods for solving these problems. One of these methods is to take subsamples of the nonrespondents. Effort is made to obtain responses from all elements in the final subsample. This technique is called nonrespondent subsampling. In this chapter nonrespondent subsampling for unit nonresponse is investigated. Section 3.1 reviews the processes of nonrespondent subsampling. Section 3.2 presents notations used in this chapter. Section 3.3 states and proves, where necessary, various theorems used in chapter 6 in computing the parameters and their variances in nonrespondent subsampling.

3.1 Overview

When nonresponse occurs, the conventional method is to take a random subsample of the individuals who have not responded and obtain a response from everyone in this subsample. This approach was introduced by *Hansen & Hurwitz* (1946) for

the survey in which mail survey was used for the first attempt and at the second stage personal interviews nonrespondents of the first stages were conducted. Their procedure is applicable to other types of interviews, e.g., telephone followed by personal interviews. In the *Hansen & Hurwitz* approach, the population of size N is assumed to be composed of two artificial strata of sizes N_1 and $N_2 = N - N_1$, of "respondents" and "nonrespondents". The initial simple random sample of size n results in n_{11} respondents and n_{12} nonrespondents. A subsample of size $n'_{12} = \frac{n_{12}}{k}$, where k is a constant predetermined by the experience of the survey designer, is drawn from the n_{12} nonrespondents. Response is assumed to be obtained from all of the n'_{12} units. In practice, it may not be possible to obtain information from all of the n'_{12} units and some adjustments to the estimates have to be made, accounting for the "hard-core" nonrespondents.

Hansen & Hurwitz (1946) also give optimum values for n and k which minimise the expected cost C where $C = C_0 n + C_1 n_{11} + C_2 n'_{12}$, C_0 is the initial cost of "setting up" the survey, C_1 is the cost per unit of obtaining the responses from the n_{11} units and processing them, and C_2 is the cost per unit of contacting the subsampled units and of obtaining and processing responses from them. Usually C_2 is much larger than C_1, since additional effort is needed for contacting the nonrespondents and eliciting response from them. Optimum values k and n for minimising the expected cost for a prescribed variance $V = (\frac{N-n}{Nn})S^2 + \frac{(k-1)W_{12}S_{12}^2}{n} = \epsilon^2$ are given by

$$k_{opt} = [\frac{C_2(S^2 - W_{12}S_{12}^2)}{S_{12}^2(C_0 + C_1 W_{11})}]^{1/2},$$

and

$$n_{opt} = n_0[1 + \frac{(k_{opt}-1)W_{12}S_{12}^2}{S^2}],$$

where W_{11} and W_{12} are the population proportion for response and nonresponse stratum respectively, S_{12}^2 and S^2 are the population variance for response stratum

and for the whole population, and $n_0 = \frac{NS^2}{NV+S^2}$ is the sample size required to achieve a variance V if there is no nonresponse.

Srinath (1971) suggests a refinement to the above: the size of subsample is $n_{12}^{*'} = \frac{n_{12}^2}{k^*n+n_{12}} = \frac{nw_{12}^2}{k^*+w_{12}} = \frac{n_{12}^{'}kw_{12}}{k^*+w_{12}}$, where k^* is predetermined by the experience of the survey designer. Under *Srinath's* alternative rule, optimum values of k^* and n^* which minimise the expected cost $C = C_0n + C_1n_{11} + C_2n_{12}^{'*}$ for the same fixed variance $V = \epsilon^2$ as for the *Hansen & Hurwitz* method, are given by

$$k^* = (k_{opt} - 1)W_{12},$$

and

$$n_{opt}^* = n_0(1 + k_{opt}^* \frac{S_{12}^2}{S^2}).$$

Solution for both k and k^* require a knowledge of nonresponse population proportion W_{12}. This is often estimated from experience. *Srinath* (1971) shows that if W_{12} is known, then his rule gives the same expected cost for a desired precision as *Hansen & Hurwitz's* . However, if W_{12} is not known accurately, then the subsampling rule suggested by *Srinath* will adjust the sample size in order to maintain the predetermined variance at a slightly increased cost. Moreover note that *Srinath* (1971) uses less nonrespondent sampling and hence comparisons should be done cautiously.

There are several extensions that broaden the applicability of nonrespondent subsampling and some of these are briefly outline below.

El-Badry (1956) extended the *Hansen & Hurwitz* procedure to more than two attempts. These attempts are made through follow-up letters or calls before a nonrespondent subsample for personal interviews is chosen. Complete response is assumed in the personal interviews. Data obtained from each attempt and the personal interviews are combined to produce the final estimate.

Rao (1968) considers one subsampling scheme but where the list from which the sample is drawn contains an unknown number of duplicated units.

Rao (1973) shows how to apply post-stratification theory as a special case of *Hansen & Hurwitz* nonresponse method. This technique leads to a simple situation for the optimal design of analytical surveys involving comparison of group means when the groups are not identifiable in advance.

Rao & Hughes (1983) extend the above theory to give an optimised solution for nonrespondent subsampling applied to mail surveys when the object is to estimate the difference in means between two domains that may not be defined by strata used for sampling. Two alternative sampling schemes are considered and in each case, nonrespondents are subsampled twice.

Singh & Sedransk (1978) apply Bayesian nonrespondent subsampling as a special case of estimation of finite population parameters when there is nonresponse.

Singh (1983) presents an alternative model for Bayesian nonrespondent subsampling for population mean, proportion and regression coefficients.

Rao & Ghangurde (1972) consider a finite Bayesian population mean in two schemes of two-phase sample design with nonresponse problem.

Ericson (1967) presents the results pertaining to estimation of the population mean, which assumes that strata variances are known, when there is nonresponse.

For a general nonrespondent subsampling scheme, the subsampling can be carried to several stages. The procedure can be described as follow:

In simple random sampling with or without replacement, an initial sample of size n is chosen from a population N. Let n_{11} and n_{12} be the respondents size

and nonrespondents size respectively. If a subsample of size $n'_{12} = \frac{n_{12}}{k_1}$, where k_1 is a predetermined value, is chosen from the nonresponse group in the initial sample, and if there is no nonresponse in this stage, this scheme is called a one-subsampling scheme. However, if there are still some nonrespondents and a second subsampling of size $n'_{22} = \frac{n_{22}}{k_2}$, where k_2 is determined in advance, is conducted and at this stage if there is no nonresponse, this is called a two-subsampling scheme. If there are some nonrespondents in the second stage, nonrespondent subsampling will continue until the L^{th} stage assume that there is no nonresponse, which is called a L-subsampling scheme.

In this thesis, non-Bayesian nonrespondent subsampling is studied for one-subsampling or two-subsampling schemes. The relevant theorems follow and see chapter 6 for final results.

3.2 Notation

The following notation used through this chapter is

n is an initial random sample size.

n_{ij} is a size for initial sample ($i = 1$), first subsampling ($i = 2$), response ($j = 1$) and nonresponse ($j = 2$), where $n = n_{11} + n_{12}$.

$n'_{i2} = \frac{n_{i2}}{k_i}$ is the size of i^{th} nonrespondent subsampling, for $i = 1, 2$.

y is a sum of characteristics of interest in sample size n.

y_{ij} is a sum of characteristics of interest for initial sample ($i = 1$), first subsampling ($i = 2$), response ($j = 1$) and nonresponse ($j = 2$), where $y = y_{11} + y_{12}$.

y'_{ij} is a sum of characteristics of interest of i^{th} nonrespondent subsampling, for $i = 1, 2$, with response group ($j = 1$) and nonresponse group ($j = 2$), and $y'_{12} = y_{21} + y_{22}$.

$\hat{\mu}_{ij}$ is a mean of characteristics of interest for initial sample ($i = 1$), first subsampling ($i = 2$), response ($j = 1$) and nonresponse ($j = 2$).

$\hat{\mu}'_{ij}$ is a mean of characteristics of interest of i^{th} nonrespondent subsampling, for $i = 1, 2$, with response group ($j = 1$) and nonresponse group ($j = 2$).

k_i is a predetermined value to get a size for i^{th} nonrespondent subsampling, for $i = 1, 2$.

S^2 or σ^2 is a population variance.

S^2_{i2} or σ^2_{i2} is a population variance for a first nonrespondent subsampling ($i = 1$) and a second nonrespondent subsampling ($i = 2$).

s^2_m is a sample response variance where $m = n_{11} + n_{21} + n'_{22}$.

s^2_{i2} is a sample response variance in a first nonrespondent subsampling ($i = 1$) and a second nonrespondent subsampling ($i = 2$).

3.3 Nonrespondent Subsampling Theory

This section proves theorems relating to nonrespondent subsampling. Theorems 3.1-3.8 discuss nonrespondent subsampling in equal probability sampling for estimator of the population mean. Theorems 3.9-3.12 discuss nonrespondent subsampling in unequal probability sampling when estimating the population total. These theorems will be proved for two-subsampling schemes. For one-subsampling schemes, proofs follow in a similar way. These theorems parallel the basic sampling theorems 2.1-2.10 in section 2.3.3.

These theorems about the estimated variance in this chapter for simple random sampling replace σ^2 in the case of sampling with replacement or S^2 in the case of sampling without replacement by using response sample variance in lemma 3.1 defined below. This is also the case in stratified and post-stratified random sampling.

Before theorems with nonrespondent subsampling are introduced, the following two lemmas, which are helpful to prove theorem 3.1-3.12, are stated without proof. These two lemmas are from *Cochran* (1977).

Lemma 3.1 *Sample Variance*

1) s_m^2 is an estimator of a population variance S^2 (or σ^2), where m is the response size of combination with the initial plan, first subsampling and second subsampling.

2) s_{12}^2 is an unbiased estimator of a population variance in a first nonresponse stratum S_{12}^2 (or σ_{12}^2).

3) s_{22}^2 is an unbiased estimator of a population variance in a second nonresponse stratum S_{22}^2 (or σ_{22}^2).

■

Lemma 3.2 *Expectation and Variance in Three-Phase Sampling*

A expectation of estimator $\hat{\theta}$ in three-phase sampling is given by

$$E(\hat{\theta}) = E_1 E_2 E_3(\hat{\theta}),$$

where E_1, E_2 and E_3 are the expectation of parameter θ for the initial attempt, first and second nonrespondent subsampling respectively.

A variance of estimator $\hat{\theta}$ in three-phase sampling is given by

$$V(\hat{\theta}) = V_1 E_2 E_3(\hat{\theta}) + E_1 V_2 E_3(\hat{\theta}) + E_1 E_2 V_3(\hat{\theta}),$$

where the variances are of the estimator of theta, V_1, V_2 and V_3, are the variance of parameter $\hat{\theta}$ for the initial attempt, first and second nonrespondent subsampling respectively.

■

Hansen & Hurwitz (1946), *Srinath* (1971) and *Cochran* (1977) give results without proof for one-subsampling scheme in simple random sampling without replacement. *Rao* (1983) gives results with proof for an estimator of a population mean with one-subsampling scheme in simple random sampling without replacement. *El-Badry* (1956) and *Srinath* (1971) give results without proof for multi-subsampling scheme in simple random sampling without replacement. Theorem 3.1 is proven for nonrespondent two-subsampling scheme in simple random sampling without replacement. I extend this procedure into simple random sampling with replacement as shown in theorem 3.2.

Theorem 3.1 *Simple Random Sampling without Replacement*

In simple random sampling of size n with nonrespondent two-subsampling scheme, an unbiased estimator of μ is

$$\hat{\mu}_{srs} = \frac{1}{n}(y_{11} + k_1 y_{21} + k_1 k_2 y_{22}^{'}), \tag{3.1}$$

with a variance of

$$V(\hat{\mu}_{srs}) = \frac{N-n}{nN}S^2 + \frac{1}{n}(k_1 - 1)W_{12}S_{12}^2 + \frac{1}{n}k_1(k_2 - 1)W_{22}S_{22}^2. \tag{3.2}$$

Proof: With $k_j = \frac{n_{j2}}{n_{j2}^{'}}$ for $j = 1, 2$ are constants greater than 1,

$$\hat{\mu}_{srs} = \tfrac{1}{n}[y_{11} + \tfrac{n_{12}}{n_{12}^{'}}(y_{21} + \tfrac{n_{22}}{n_{22}^{'}}y_{22}^{'})].$$

The expectation of mean estimator is given by

$$\begin{aligned}
E(\hat{\mu}_{srs}) &= E_1E_2E_3\frac{1}{n}[y_{11} + \frac{n_{12}}{n_{12}^{'}}(y_{21} + \frac{n_{22}}{n_{22}^{'}}y_{22}^{'})] \\
&= E_1E_2\frac{1}{n}[y_{11} + \frac{n_{12}}{n_{12}^{'}}(y_{21} + n_{22}E_3(\bar{y}_{22}^{'}))] \\
&= E_1E_2\frac{1}{n}[y_{11} + \frac{n_{12}}{n_{12}^{'}}(y_{21} + n_{22}\bar{y}_{22})] \\
&= E_1\frac{1}{n}[y_{11} + n_{12}E_2(\bar{y}_{12}^{'})]
\end{aligned}$$

$$
\begin{aligned}
&= E_1\frac{1}{n}[y_{11}+y_{12}]\\
&= E_1(\hat{\mu})\\
&= \mu.
\end{aligned}
$$

The results that $E_3[y_{i1}] = y_{i1}$ for $i = 1, 2$ and $E_3[\bar{y}_{22}^{'}] = \bar{y}_{22}$ and similarly $E_2[y_{11}] = y_{11}$ and $E_2[\bar{y}_{12}^{'}] = \bar{y}_{12}$ and also $y_{21} + n_{22}\bar{y}_{22} = y_{12}^{'}$ and similarly $y_{11} + n_{12}\bar{y}_{12} = \sum y$ above have been used.

To prove equation 3.2 note that

$$
\begin{aligned}
V_1E_2E_3(\hat{\mu}_{srs}) &= V_1(\hat{\mu})\\
&= \frac{N-n}{nN}S^2,\\
E_1V_2E_3(\hat{\mu}_{srs}) &= E_1(\frac{n_{12}}{n})^2V_2(\bar{y}_{12}^{'})\\
&= E_1(\frac{n_{12}}{n})^2\frac{n_{12}-n_{12}^{'}}{n_{12}^{'}n_{12}}S_{12}^2\\
&= E_1[\frac{w_{12}}{n}(k_1-1)S_{12}^2]\\
&= \frac{1}{n}(k_1-1)W_{12}S_{12}^2,
\end{aligned}
$$

and

$$
\begin{aligned}
E_1E_2V_3(\hat{\mu}_{srs}) &= E_1E_2V_3\frac{1}{n}[y_{11}+\frac{n_{12}}{n_{12}^{'}}y_{21}+\frac{n_{12}n_{22}}{n_{12}^{'}n_{22}^{'}}y_{22}^{'}]\\
&= E_1E_2\frac{1}{n^2}k_1^2n_{22}^2V_3(\bar{y}_{22}^{'})\\
&= \frac{k_1^2}{n^2}E_1E_2n_{22}^2(\frac{n_{22}-n_{22}^{'}}{n_{22}^{'}n_{22}})S_{22}^2\\
&= \frac{k_1^2}{n^2}E_1E_2n_{22}(k_2-1)S_{22}^2\\
&= \frac{k_1^2(k_2-1)}{n^2}S_{22}^2E_1[n_{12}^{'}\frac{N_{22}}{N_{12}}]\\
&= \frac{k_1^2(k_2-1)}{n^2}S_{22}^2E_1[\frac{n_{12}N_{22}}{k_1N_{12}}]\\
&= \frac{k_1(k_2-1)}{n^2}S_{22}^2\frac{W_{22}}{W_{12}}nW_{12}\\
&= \frac{k_1(k_2-1)}{n}W_{22}S_{22}^2.
\end{aligned}
$$

The results that $w_{12} = \frac{n_{12}}{n}$, $E_1(w_{12}) = W_{12}$ and $E_2(n_{22}) = n'_{12}\frac{N_{22}}{N_{12}}$ above have been used.

Equation 3.2 follows by lemma 3.2.

■

Simple random sampling with replacement can be applied as *Srinath* (1971) suggested for *SRSWOR*. An unbiased population mean estimator under this sampling plan is the same as *SRSWOR* but its variance is a little different depending on the sampling procedure.

Theorem 3.2 *Simple Random Sampling with Replacement*

In simple random sampling of size n with nonrespondent two-subsampling scheme, an unbiased estimator of μ is

$$\hat{\mu}_{srs} = \frac{1}{n}(y_{11} + k_1 y_{21} + k_1 k_2 y'_{22}). \tag{3.3}$$

with a variance of

$$V(\hat{\mu}_{srs}) = \frac{\sigma^2}{n} + \frac{k_1}{n} W_{12}\sigma^2_{12} + \frac{k_1 k_2}{n} W_{22}\sigma^2_{22}. \tag{3.4}$$

Proof: An unbiased estimator of the population mean is proved as in theorem 3.1. To prove equation 3.4 note that

$$\begin{aligned}
V_1E_2E_3(\hat{\mu}_{srs}) &= V_1(\hat{\mu}) \\
&= \frac{\sigma^2}{n}, \\
E_1V_2E_3(\hat{\mu}_{srs}) &= E_1(\frac{n_{12}}{n})^2 V_2(\bar{y}'_{12}) \\
&= E_1(\frac{n_{12}}{n})^2 \frac{\sigma^2_{12}}{n'_{12}} \\
&= k_1 \frac{\sigma^2_{12}}{n^2} E_1(n_{12}) \\
&= \frac{k_1}{n} W_{12}\sigma^2_{12},
\end{aligned}$$

and

$$E_1E_2V_3(\hat{\mu}_{srs}) = E_1E_2V_3\frac{1}{n}[y_{11} + \frac{n_{12}}{n'_{12}} y_{21} + \frac{n_{12}n_{22}}{n'_{12}n'_{22}} y'_{22}]$$

$$
\begin{aligned}
&= E_1E_2\frac{1}{n^2}(\frac{n_{12}n_{22}}{n_{12}^{'}})^2V_3(\bar{y}_{22}^{'}) \\
&= E_1E_2\frac{1}{n^2}(\frac{n_{12}n_{22}}{n_{12}^{'}})^2\frac{\sigma_{22}^2}{n_{22}^{'}} \\
&= (\frac{k_1}{n})^2k_2\sigma_{22}^2E_1E_2(n_{22}) \\
&= (\frac{k_1}{n})^2k_2\sigma_{22}^2E_1(\frac{n_{12}W_{22}}{k_1W_{12}}) \\
&= \frac{k_1k_2}{n^2}\sigma_{22}^2\frac{W_{22}}{W_{12}}E_1(n_{12}) \\
&= \frac{k_1k_2}{n}W_{22}\sigma_{22}^2.
\end{aligned}
$$

The results that $E_2(n_{22}) = n_{12}^{'}\frac{N_{22}}{N_{12}}$ and $E_1(n_{12}) = nW_{12}$ above have been used.

Equation 3.4 follows by lemma 3.2.

■

Let us assume that the population consists of H strata of sizes N_h for $h=1,...,H$ In the h^{th} stratum, let N_{h1} be the size of the respondents and $N_{h2} = N_h - N_{h1}$ is the size of the nonrespondents. The initial simple random sample of size n_h from the h^{th} stratum results in n_{h11} respondents and n_{h12} nonrespondents. The first subsampling with size $n_{h12}^{'} = \frac{n_{h12}}{k_{h1}}$, where k_{h1} is fixed in advance, is drawn randomly. If there is no nonresponse in this stage, this scheme is called nonrespondent one-subsampling. *Rao* (1983) gives the mean estimator results with proof for nonrespondent one-subsampling in stratified random sampling without replacement. I extend this idea. Assume that if the response is obtained on part of subsample, n_{h21} units, the second subsampling with size $n_{h22}^{'} = \frac{n_{h22}}{k_{h2}}$, where n_{h22} are the nonrespondents from the first subsampling and k_{h2} is fixed in advance, is drawn randomly and response is assumed completely. This idea leads to theorem 3.3.

Theorem 3.3 *Stratified Random Sampling without Replacement*

In stratified random sampling of size n_h with $\sum_{h=1}^{H} n_h = n$ and nonrespondent two-subsampling scheme, an unbiased estimator of μ is

$$\hat{\mu}_{st} = \sum_{h=1}^{H}\frac{W_h}{n_h}(y_{h11} + k_{h1}y_{h21} + k_{h1}k_{h2}y_{h22}^{'}) = \sum_{h=1}^{H}W_h\hat{\mu}_{h,srs}, \tag{3.5}$$

where $W_h = \frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{N_h - n_h}{n_h N_h} S_h^2 + \sum_{h=1}^{H} \frac{W_h^2}{n_h}(k_{h1}-1)W_{h12}S_{h12}^2 + \sum_{h=1}^{H} \frac{W_h^2}{n_h} k_{h1}(k_{h2}-1)W_{h22}S_{h22}^2. \tag{3.6}$$

Proof: By theorem 3.1, for each stratum $E(\hat{\mu}_{h,srs}) = \mu_h$. Hence

$$\begin{aligned} E(\hat{\mu}_{st}) &= \sum_{h=1}^{H} W_h E(\hat{\mu}_{h,srs}) \\ &= \sum_{h=1}^{H} \frac{N_h}{N} \mu_h \\ &= \mu. \end{aligned}$$

Since the selections in different strata are independent,

$$V(\hat{\mu}_{st}) = \textstyle\sum_{h=1}^{H} (W_h)^2 V(\hat{\mu}_{h,srs}).$$

Equation 3.6 follows from theorem 3.1.

■

Stratified random sampling with replacement can be applied as *Rao* (1983) suggested for *STWOR*. An unbiased population mean estimator under this sampling plan is the same as *STWOR* but its variance can be different depending on the sampling procedure.

Theorem 3.4 *Stratified Random Sampling with Replacement*

In stratified random sampling of size n_h with $\sum_{h=1}^{H} n_h = n$ and nonrespondent two-subsampling scheme, an unbiased estimator of μ is

$$\hat{\mu}_{st} = \sum_{h=1}^{H} \frac{W_h}{n_h}(y_{h11} + k_{h1}y_{h21} + k_{h1}k_{h2}y'_{h22}) = \sum_{h=1}^{H} W_h \hat{\mu}_{h,srs}, \tag{3.7}$$

where $W_h = \frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{st}) = \sum_{h=1}^{H} W_h^2 \frac{\sigma_h^2}{n_h} + \sum_{h=1}^{H} \frac{W_h^2}{n_h} k_{h1} W_{h12} \sigma_{h12}^2 + \sum_{h=1}^{H} \frac{W_h^2}{n_h} k_{h1}k_{h2} W_{h22} \sigma_{h22}^2. \tag{3.8}$$

Proof: Proof for unbiasedness of the mean estimator and for its variance is similar to that for theorem 3.3.

■

I extend the idea of nonrespondent two-subsampling into post-stratified random sampling design with and without replacement schemes. This leads to theorems 3.5-3.8.

In complex large-scale survey, post-stratification is a common technique used for improving the efficiency of estimators. It should improve efficiency when the auxiliary variable is chosen correctly. Values of the auxiliary variables for establishments (e.g., business size, type of establishment, form of legal organisation, and other economic factors) are sometimes unavailable for sample design at individual level. A census of industry may, however, provide aggregate information on such variables that can be used at the estimation stage. After sample selection, sampled auxiliary variables are used to stratify the sample (post-stratification) and the known number of units in the h^{th} post-stratum from the aggregate information, N_h, is used as a weight to estimate the h^{th} post-stratum total. The nonresponse units are subsampled to collect data and two scenarios, *SCENARIO:A* and *SCENARIO:B*, will be considered. The *SCENARIO:A* is the scheme when auxiliary variables about nonrespondents are available for post-stratification. When there is no auxiliary variable about nonrespondents for post-stratification it is called *SCENARIO:B*.

For *SCENARIO:A*, simple random sample of n units is selected from the entire population. After conducting a survey, n_{11} respondents can be classified into H_s post-strata with n_{h11} respondents and $\sum_{h=1}^{H_s} n_{h11} = n_{11}$. The $n_{12} = n - n_{11}$ nonrespondents also have enough information to be classified into H_s post-strata with size n_{h12} units in post-stratum h. In each post-stratum, subsampling on nonrespondents are conducted with size $n'_{h12} = \frac{n_{h12}}{k_{h1}}$, where $k_{h1} > 0$ is fixed in advance and n_{h21} units are assumed respond. The nonrespondents at this stage are then sub-

jected to a second subsampling of size $n'_{h22} = \frac{n_{h22}}{k_{h2}}, k_{h2} > 0$ is predetermined and $n_{h22} = n'_{h12} - n_{h21}$. A total response is assumed from this stage. The mean and variance of mean estimator under this scenario are given below in theorem 3.5 and 3.6.

Theorem 3.5 *Post-stratified Random Sampling without Replacement*

In post-stratified random sampling with nonrespondent two-subsampling scheme: scenario A, an unbiased estimator of μ is

$$\hat{\mu}_{pt} = \sum_{h=1}^{H_s} \frac{W_h}{n_h}(y_{h11} + k_{h1}y_{h21} + k_{h1}k_{h2}y'_{h22}) = \sum_{h=1}^{H_s} W_h\hat{\mu}_{h,srs}, \tag{3.9}$$

where $W_h = \frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{pt}) \approx \tfrac{N-n}{nN}\textstyle\sum_{h=1}^{H_s} W_hS_h^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{H_s}(1-W_h)S_h^2 +$$

$$\sum_{h=1}^{H_s} \frac{W_h^2}{n_h}(k_{h1}-1)W_{h12}S_{h12}^2 + \sum_{h=1}^{H_s} \frac{W_h^2}{n_h}k_{h1}(k_{h2}-1)W_{h22}S_{h22}^2. \tag{3.10}$$

Proof: Proof for unbiasedness of the mean estimator is similar to that for theorem 3.3.

To prove equation 3.10 note that the variance of $\hat{\mu}_{pt}$ in the initial sample phase follows as in theorem 2.6:

$$\begin{aligned} V_1E_2E_3(\hat{\mu}_{pt}) &= V_1(\sum_{h=1}^{Hs} W_h\hat{\mu}_h) \\ &\approx \frac{N-n}{nN}\sum_{h=1}^{Hs} W_hS_h^2 + \frac{N-n}{n^2N}\sum_{h=1}^{Hs}(1-W_h)S_h^2. \end{aligned}$$

The variance of $\hat{\mu}_{pt}$ in the first and second subsampling phase can be used as in theorem 3.3 for stratified random sampling scheme:

$$\begin{aligned} E_1V_2E_3(\hat{\mu}_{pt}) &= E_1V_2\sum_{h=1}^{Hs}\frac{W_h}{n_h}[y_{h11} + n_{h12}\bar{y}'_{h12}] \\ &= \sum_{h=1}^{Hs}\frac{W_h^2}{n_h}(k_{h1}-1)W_{h12}S_{h12}^2, \end{aligned}$$

where $E_1(w_{h12}) = W_{h12}$ and

$$\begin{aligned} E_1E_2V_3(\hat{\mu}_{pt}) &= E_1E_2V_3[\sum_{h=1}^{Hs}\frac{W_h}{n_h}(y_{h11}+\frac{n_{h12}}{n'_{h12}}y_{h21}+\frac{n_{h12}n_{h22}}{n'_{h12}n'_{h22}}y'_{h22})] \\ &= \sum_{h=1}^{Hs}[\frac{W_h^2}{n_h}k_{h1}(k_{h2}-1)W_{h22}S_{h22}^2]. \end{aligned}$$

Equation 3.10 follows by lemma 3.2.

■

Theorem 3.6 *Post-stratified Random Sampling with Replacement*

In post-stratified random sampling with nonrespondent two-subsampling scheme: scenario A, an unbiased estimator of μ is

$$\hat{\mu}_{pt} = \sum_{h=1}^{H_s}\frac{W_h}{n_h}(y_{h11}+k_{h1}y_{h21}+k_{h1}k_{h2}y'_{h22}) = \sum_{h=1}^{H_s}W_h\hat{\mu}_{h,srs}, \tag{3.11}$$

where $W_h = \frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{pt}) \approx \tfrac{1}{n}\textstyle\sum_{h=1}^{H_s}W_h\sigma_h^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{H_s}(1-W_h)\sigma_h^2 +$$

$$\sum_{h=1}^{H_s}\frac{W_h^2}{n_h}k_{h1}W_{h12}\sigma_{h12}^2 + \sum_{h=1}^{H_s}\frac{W_h^2}{n_h}k_{h1}k_{h2}W_{h22}\sigma_{h22}^2. \tag{3.12}$$

Proof: Proof for unbiasedness of the mean estimator is similar to that for theorem 3.3.

To prove equation 3.12 note that the variance of $\hat{\mu}_{pt}$ in the initial sample phase follows as in theorem 2.5:

$$\begin{aligned} V_1E_2E_3(\hat{\mu}_{pt}) &= V_1(\sum_{h=1}^{Hs}W_h\hat{\mu}_h) \\ &\approx \frac{1}{n}\sum_{h=1}^{Hs}W_h\sigma_h^2 + \frac{1-f}{n^2}\sum_{h=1}^{Hs}(1-W_h)\sigma_h^2, \end{aligned}$$

where $f = \frac{n}{N}$.

The variance of $\hat{\mu}_{pt}$ in the first and second subsampling phase can be used as in theorem 3.4 for stratified random sampling scheme:

$$\begin{aligned} E_1V_2E_3(\hat{\mu}_{pt}) &= E_1V_2\sum_{h=1}^{Hs}\frac{W_h}{n_h}[y_{h11}+n_{h12}\bar{y}_{h12}^{'}] \\ &= \sum_{h=1}^{Hs}\frac{W_h^2}{n_h}k_{h1}W_{h12}\sigma_{h12}^2, \end{aligned}$$

and

$$\begin{aligned} E_1E_2V_3(\hat{\mu}_{pt}) &= E_1E_2V_3[\sum_{h=1}^{Hs}\frac{W_h}{n_h}(y_{h11}+\frac{n_{h12}}{n_{h12}^{'}}y_{h21}+\frac{n_{h12}n_{h22}}{n_{h12}^{'}n_{h22}^{'}}y_{h22})] \\ &= E_1E_2\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}(\frac{n_{h12}}{n_{h22}^{'}})^2n_{h22}^2V_3(\bar{y}_{h22}^{'})] \\ &= \sum_{h=1}^{Hs}\frac{W_h^2}{n_h}k_{h1}k_{h2}W_{h22}\sigma_{h22}^2. \end{aligned}$$

Equation 3.12 follows by lemma 3.2.

■

For *SCENARIO:B*, simple random sample of n units is selected from the entire population. After conducting a survey, n_{11} respondents can be classified into H_s post-strata with n_{h11} respondents and $\sum_{h=1}^{H_s}n_{h11}=n_{11}$. There is no information for $n_{12}=n-n_{11}$ nonrespondents. Subsampling on nonrespondents will be conducted with size $n_{12}^{'}=\frac{n_{11}}{k_1}$ units, where $k_1>0$ is predetermined, and n_{h21} units are assumed respond in the post-stratum h. The nonrespondents at this stage are then subjected to a second subsampling of size $n_{22}^{'}=\frac{n_{22}}{k_2}$ units, where $k_2>0$ is predetermined and $n_{22}=n_{12}^{'}-n_{21}$, and a total response is assumed in this stage. These response units can be classified into H_s post-strata with $n_{h22}^{'}$ respondents and $\sum_{h=1}^{H_s}n_{h22}^{'}=n_{22}$. The mean and variance of mean estimator under this scenario are given below in theorem 3.7 and 3.8.

Theorem 3.7 *Post-stratified Random Sampling without Replacement*

In post-stratified random sampling with nonrespondent two-subsampling scheme: scenario B, an unbiased estimator of μ is

$$\hat{\mu}_{pt}=\sum_{h=1}^{H_s}\frac{W_h}{n_h}(y_{h11}+k_{h1}y_{h21}+k_{h1}k_{h2}y_{h22}^{'})=\sum_{h=1}^{H_s}W_h\hat{\mu}_{h,srs}, \tag{3.13}$$

where $W_h=\frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{pt}) \approx \tfrac{N-n}{nN}\textstyle\sum_{h=1}^{H_s} W_h S_h^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{H_s}(1-W_h)S_h^2 +$$

$$\tfrac{1}{n}\textstyle\sum_{h=1}^{H_s} W_h W_{h12}(k_{h1}-1)S_{h12}^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{Hs} W_{h12}S_{h12}^2(k_{h1}-1)(1-W_h) +$$

$$\frac{1}{n}\sum_{h=1}^{H_s} W_h W_{h22}k_{h1}(k_{h2}-1)S_{h22}^2 + \frac{N-n}{n^2N}\sum_{h=1}^{H_s}(1-W_h)W_{h22}k_{h1}(k_{h2}-1)S_{h22}^2. \quad (3.14)$$

Proof: Proof for unbiasedness of the mean estimator is similar to that for theorem 3.3.

To prove equation 3.14 note that the variance of $\hat{\mu}_{pt}$ in the initial sample phase follows as in theorem 2.6:

$$\begin{aligned} V_1E_2E_3(\hat{\mu}_{pt}) &= V_1(\sum_{h=1}^{Hs} W_h\hat{\mu}_h) \\ &\approx \frac{N-n}{nN}\sum_{h=1}^{Hs} W_hS_h^2 + \frac{N-n}{n^2N}\sum_{h=1}^{Hs}(1-W_h)S_h^2. \end{aligned}$$

Similarly for the variance of $\hat{\mu}_{pt}$ in the first and second subsampling phase, the variance of $\hat{\mu}_{pt}$ follows as in theorem 2.6:

$$\begin{aligned} E_1V_2E_3(\hat{\mu}_{pt}) &= E_1V_2\sum_{h=1}^{Hs}\frac{W_h}{n_h}[y_{h11}+n_{h12}\bar{y}'_{h12}] \\ &= E_1\sum_{h=1}^{Hs}\frac{W_h^2}{n_h^2}[V_2(y_{h1})+n_{h12}^2V_2(\bar{y}'_{h12})] \\ &= E_1\sum_{h=1}^{Hs}\frac{W_h^2}{n_h^2}[n_{h12}^2(\frac{n_{h12}-n'_{h12}}{n'_{h12}n_{h12}})S_{h12}^2] \\ &= E\sum_{h=1}^{Hs}\frac{W_h^2}{n_h^2}S_{h12}^2(k_{h1}-1)E_s(n_{h12}|S) \\ &= E\sum_{h=1}^{Hs}\frac{W_h^2}{n_h}S_{h12}^2(k_{h1}-1)W_{h12} \\ &\approx \sum_{h=1}^{Hs} W_h^2S_{h12}^2(k_{h1}-1)W_{h12}[\frac{1}{nW_h}+\frac{(1-f)(1-W_h)}{n^2W_h^2}] \\ &= \frac{1}{n}\sum_{h=1}^{Hs} W_hW_{h12}(k_{h1}-1)S_{h12}^2 + \frac{N-n}{n^2N}\sum_{h=1}^{Hs}(1-W_h)W_{h12}(k_{h1}-1)S_{h12}^2, \end{aligned}$$

where $E_s(n_{h12}|S) = n_hW_{h12}$ and $f = \frac{n}{N}$.

$$E_1E_2V_3(\hat{\mu}_{pt}) = E_1E_2V_3[\sum_{h=1}^{Hs}\frac{W_h}{n_h}(y_{h11}+\frac{n_{h12}}{n'_{h12}}y_{h21}+\frac{n_{h12}n_{h22}}{n'_{h12}n'_{h12}}y'_{h22})]$$

$$
\begin{aligned}
&= E_1E_2\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}(\frac{n_{h12}}{n_{h12}^{'}})^2 n_{h22}^2 V_3(\bar{y}_{h22}^{'})] \\
&= E_1E_2\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}k_{h1}^2 n_{h22}^2(\frac{n_{h22}-n_{h22}^{'}}{n_{h22}^{'}n_{h22}})S_{h22}^2] \\
&= E_1E_2\sum_{h=1}^{Hs}\frac{W_h^2}{n_h^2}k_{h1}^2[n_{h22}(k_{h2}-1)]S_{h22}^2 \\
&= E_1\sum_{h=1}^{Hs}\frac{W_h^2}{n_h^2}k_{h1}^2(k_{h2}-1)S_{h22}^2E_2(n_{h22}) \\
&= E_1\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}k_{h1}^2(k_{h2}-1)S_{h22}^2(n_{h12}^{'}\frac{N_{h22}}{N_{h12}})] \\
&= E\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}k_{h1}^2(k_{h2}-1)S_{h22}^2\frac{W_{h22}}{W_{h12}}E_s(n_{h12}^{'}|S)] \\
&= E\sum_{h=1}^{Hs}[\frac{W_h^2}{n_h^2}k_{h1}^2(k_{h2}-1)S_{h22}^2\frac{W_{h22}n_hW_{h12}}{W_{h12}k_{h1}}] \\
&= \sum_{h=1}^{Hs}[W_h^2k_{h1}(k_{h2}-1)S_{h22}^2W_{h22}E(\frac{1}{n_h})] \\
&\approx \sum_{h=1}^{Hs}W_h^2k_{h1}(k_{h2}-1)S_{h22}^2W_{h22}[\frac{1}{nW_h}+\frac{(1-f)(1-W_h)}{n^2W_h^2}]
\end{aligned}
$$

$$= \tfrac{1}{n}\textstyle\sum_{h=1}^{Hs}W_hk_{h1}(k_{h2}-1)W_{h22}S_{h22}^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{Hs}(1-W_h)k_{h1}(k_{h2}-1)W_{h22}S_{h22}^2,$$

where $E_2(n_{h22}) = n_{h12}^{'}\frac{N_{h22}}{N_{h12}}$ and $f = \frac{n}{N}$.

Equation 3.14 follows by lemma 3.2.

■

Theorem 3.8 *Post-stratified Random Sampling with Replacement*

In post-stratified random sampling with nonrespondent two-subsampling scheme: scenario B, an unbiased estimator of μ is

$$\hat{\mu}_{pt} = \sum_{h=1}^{H_s}\frac{W_h}{n_h}(y_{h11}+k_{h1}y_{h21}+k_{h1}k_{h2}y_{h22}^{'}) = \sum_{h=1}^{H}W_h\hat{\mu}_{h,srs}, \tag{3.15}$$

where $W_h = \frac{N_h}{N}$, with a variance of

$$V(\hat{\mu}_{pt}) \approx \tfrac{1}{n}\textstyle\sum_{h=1}^{H_s}W_h\sigma_h^2 + \tfrac{N-n}{n^2N}\sum_{h=1}^{H_s}(1-W_h)\sigma_h^2 +$$

$$\sum_{h=1}^{H_s} \frac{1}{n} W_h W_{h12} k_{h1} \sigma_{h12}^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{H_s} (1 - W_h) W_{h12} k_{h1} \sigma_{h12}^2 +$$

$$\frac{1}{n} \sum_{h=1}^{H_s} W_h k_{h1} k_{h2} W_{h22} \sigma_{h22}^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{H_s} (1 - W_h) W_{h22} k_{h1} k_{h2} \sigma_{h22}^2. \tag{3.16}$$

Proof: Proof for unbiasedness of the mean estimator is similar to that for theorem 3.3.

To prove equation 3.16 note that the variance of $\hat{\mu}_{pt}$ in the initial sample phase follows as in theorem 2.5:

$$\begin{aligned} V_1 E_2 E_3(\hat{\mu}_{pt}) &= V_1(\sum_{h=1}^{Hs} W_h \hat{\mu}_h) \\ &\approx \frac{1}{n} \sum_{h=1}^{Hs} W_h \sigma_h^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{Hs} (1 - W_h) \sigma_h^2. \end{aligned}$$

Similarly for the variance of $\hat{\mu}_{pt}$ in the first and second subsampling phase, the variance of $\hat{\mu}_{pt}$ follows as in theorem 2.5:

$$\begin{aligned} E_1 V_2 E_3(\hat{\mu}_{pt}) &= E_1 V_2 \sum_{h=1}^{Hs} \frac{W_h}{n_h} [y_{h11} + n_{h12} \bar{y}'_{h12}] \\ &= E_1 \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} [V_2(y_{h11}) + n_{h12}^2 V_2(\bar{y}'_{h12})] \\ &= E_1 \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} [n_{h12}^2 \frac{\sigma_{h12}^2}{n'_{h12}}] \\ &= E \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} k_{h1} \sigma_{h12}^2 E_S(n_{h12}|S) \\ &= E \sum_{h=1}^{Hs} \frac{W_h^2}{n_h} k_{h1} W_{h12} \sigma_{h12}^2 \\ &\approx \sum_{h=1}^{Hs} W_h^2 k_{h1} W_{h12} [\frac{1}{n W_h} + \frac{(1-f)(1-W_h)}{n^2 W_h^2}] \sigma_{h12}^2 \\ &= \sum_{h=1}^{H_s} \frac{1}{n} W_h W_{h12} k_{h1} \sigma_{h12}^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{H_s} (1 - W_h) W_{h12} k_{h1} \sigma_{h12}^2, \end{aligned}$$

where $E_S(n_{h12}|S) = n_h W_{h12}$ and

$$\begin{aligned} E_1 E_2 V_3(\hat{\mu}_{pt}) &= E_1 E_2 V_3 [\sum_{h=1}^{Hs} \frac{W_h}{n_h} (y_{h11} + \frac{n_{h12}}{n'_{h12}} y_{h21} + \frac{n_{h12} n_{h22}}{n'_{h12} n'_{h12}} y'_{h22})] \\ &= E_1 E_2 \sum_{h=1}^{Hs} [\frac{W_h^2}{n_h^2} (\frac{n_{h12}}{n'_{h12}})^2 n_{h22}^2 V_3(\bar{y}'_{h22})] \end{aligned}$$

$$
\begin{aligned}
&= E_1 E_2 \sum_{h=1}^{Hs} [\frac{W_h^2}{n_h^2} k_{h1}^2 n_{h22}^2 \frac{\sigma_{h22}^2}{n_{h22}^{'}}] \\
&= E_1 \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} k_{h1}^2 k_{h2} \sigma_{h22}^2 E_2(n_{h22}) \\
&= E_1 \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} k_{h1}^2 k_{h2} \sigma_{h22}^2 n_{h12}^{'} \frac{W_{h22}}{W_{h12}} \\
&= E \sum_{h=1}^{Hs} \frac{W_h^2}{n_h^2} k_{h1}^2 k_{h2} \sigma_{h22}^2 \frac{W_{h22}}{W_{h12}} E_S(\frac{n_{h12}}{k_{h1}} | S) \\
&= E \sum_{h=1}^{Hs} \frac{W_h^2}{n_h} k_{h1} k_{h2} W_{h22} \sigma_{h22}^2 \\
&\approx \sum_{h=1}^{H_s} \frac{1}{n} W_h k_{h1} k_{h2} W_{h22} \sigma_{h22}^2 + \frac{N-n}{n^2 N} \sum_{h=1}^{H_s} (1 - W_h) W_{h22} k_{h1} k_{h2} \sigma_{h22}^2,
\end{aligned}
$$

where $E_2(n_{h22}) = n_{h12}^{'} \frac{N_{h22}}{N_{h12}}$.

Equation 3.16 follows by lemma 3.2.

■

Nonrespondent subsampling can be applied to unequal probability sampling as well as to simple random sampling or stratified random sampling. *Sarndal & Swensson* (1987) shows how to use two-phase sampling with applications to nonresponse.

Sarndal et al (1992) give the results without proof for nonrespondent one-subsampling in unequal probability selection on simple random sampling without replacement but in this thesis nonrespondent two-subsampling scheme for unequal probability selection is modified in theorem 3.9. I also extend nonrespondent two-subsampling into sampling with replacement scheme and stratified random sampling design as shown in theorems 3.10-3.12.

An initial simple random sample of size n is drawn without replacement with positive inclusion probabilities π_k and π_{kl}. The survey results in n_{11} respondents and n_{12} nonrespondents. The first subsampling with size $n_{12}^{'} = \frac{n_{12}}{k_1}$, where k_1 is fixed in advance, is drawn randomly with positive conditional inclusion probabilities given partitioning into respondents and nonrespondents at the first phase. These probabilities are denoted by $\pi_{k|a_1}$ and $\pi_{kl|a_1}$. There are only n_{21} responses in this phase.

Second subsampling of size $n_{22}^{'} = \frac{n_{22}}{k_2}$ is similarly drawn, where $k_2 > 0$ is predetermined and $n_{22} = n_{12}^{'} - n_{21}$. The inclusion probability are here denoted by $\pi_{k|a_2}$ and $\pi_{kl|a_2}$ and is conditional on partitioning into respondents and nonrespondents at the second phase. This leads to theorem 3.9.

Theorem 3.9 *Random Unequal Probability Sampling without Replacement*

In a random sample of size n with nonrespondent two-subsampling, an unbiased estimator of τ is

$$\tau_{\pi ps}^{srs} = \sum_{S'} \tilde{\tilde{y}}_k = \sum_{S'} \frac{y_k}{\pi_k^*}, \tag{3.17}$$

where $S' = n_{11} + n_{21} + n_{22}^{'}$, $\tilde{\tilde{y}}_k = \frac{y_k}{\pi_k^*}$, and

$$\pi_k^* = \begin{cases} \pi_k & \text{if } k \in n_{11}, \\ \pi_k \pi_{k|a_1} & \text{if } k \in n_{21}, \\ \pi_k \pi_{k|a_1} \pi_{k|a_2} & \text{if } k \in n_{22}^{'}, \end{cases}$$

with a variance of

$$V(\hat{\tau}_{\pi ps}^{srs}) = \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} y_k y_l +$$

$$E_1 \sum_{k=1}^{n_{12}} \sum_{l=1}^{n_{12}} (\pi_{kl|a_1} - \pi_{k|a_1} \pi_{l|a_1}) \frac{y_k y_l}{\pi_k^{(1)} \pi_l^{(1)}} + E_1 E_2 \sum_{k=1}^{n_{22}} \sum_{l=1}^{n_{22}} (\pi_{kl|a_2} - \pi_{k|a_2} \pi_{l|a_2}) \frac{y_k y_l}{\pi_k^{(2)} \pi_l^{(2)}}, \tag{3.18}$$

where $\pi_k^{(1)} = \pi_k \pi_{k|a_1}$ and $\pi_k^{(2)} = \pi_k \pi_{k|a_1} \pi_{k|a_2}$.

An unbiased estimator for the variance of the sample total is

$$v(\hat{\tau}_{\pi ps}^{srs}) = \sum_{k=1}^{S'} \sum_{l=1}^{S'} \frac{\triangle_{kl}}{\pi_{kl}^*} \tilde{y}_k \tilde{y}_l +$$

$$\sum_{k=1}^{n_{21}} \sum_{l=1}^{n_{21}} \frac{\triangle_{kl|a_1}}{\pi_{kl|a_1}} \tilde{y}_k^{(1)} \tilde{y}_l^{(1)} + \sum_{k=1}^{n_{22}^{'}} \sum_{l=1}^{n_{22}^{'}} \frac{\triangle_{kl|a_2}}{\pi_{kl|a_2}} \tilde{y}_k^{(2)} \tilde{y}_l^{(2)}, \tag{3.19}$$

where $\triangle_{kl} = \pi_{kl} - \pi_k \pi_l$, $\triangle_{kl|a_1} = \pi_{kl|a_1} - \pi_{k|a_1} \pi_{l|a_1}$, $\triangle_{kl|a_2} = \pi_{kl|a_2} - \pi_{k|a_2} \pi_{l|a_2}$, $\tilde{y}_k = \frac{y_k}{\pi_k}$, $\tilde{y}_k^{(1)} = \frac{y_k}{\pi_k^{(1)}}$, $\tilde{y}_k^{(2)} = \frac{y_k}{\pi_k^{(2)}}$, and

$$\pi^*_{kl} = \begin{cases} \pi_{kl} & \text{if } k,l \in n_{11}, \\ \pi_{kl}\pi_{k|a_1} & \text{if } k \in n_{21}, l \in n_{11}, \\ \pi_{kl}\pi_{kl|a_1} & \text{if } k,l \in n_{21}, \\ \pi_{kl}\pi_{k|a_1}\pi_{k|a_2} & \text{if } k \in n'_{22}, l \in n_{11}, \\ \pi_{kl}\pi_{l|a_1}\pi_{k|a_2} & \text{if } k \in n'_{22}, l \in n_{21}, \\ \pi_{kl}\pi_{kl|a_1}\pi_{kl|a_2} & \text{if } k,l \in n'_{22}, \end{cases}$$

Proof: $\hat{\tau}^{srs}_{\pi ps}$ is unbiased estimator if $E(\hat{\tau}) = \tau$. It is noted that

$$E(\hat{\tau}^{srs}_{\pi ps}) = E_1E_2E_3(\textstyle\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{21}} \frac{y_k}{\pi_k\pi_{k|a_1}} + \sum_{k=1}^{n'_{22}} \frac{y_k}{\pi_k\pi_{k|a_1}\pi_{k|a_2}}),$$

let $c_k = \begin{cases} 1 & \text{if } y_k \in n'_{22} \\ 0 & \text{otherwise} \end{cases}$ be a *Bernoulli* random variable, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned} &= E_1E_2E_3[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{21}} \frac{y_k}{\pi_k\pi_{k|a_1}} + \sum_{k=1}^{n_{22}} \frac{c_k y_k}{\pi_k\pi_{k|a_1}\pi_{k|a_2}}] \\ &= E_1E_2[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{21}} \frac{y_k}{\pi^{(1)}_k} + \sum_{k=1}^{n_{22}} \frac{y_k}{\pi^{(1)}_k}] \\ &= E_1E_2[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n'_{12}} \frac{y_k}{\pi^{(1)}_k}], \end{aligned}$$

where $E(c_k) = \pi_{k|a_2}$, $\pi^{(1)}_k = \pi_k\pi_{k|a_1}$ and $n'_{12} = n_{21} + n_{22}$.

Let $b_k = \begin{cases} 1 & \text{if } y_k \in n'_{12} \\ 0 & \text{otherwise} \end{cases}$ be a *Bernoulli* random variable, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned} &= E_1E_2[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{12}} \frac{b_k y_k}{\pi_k\pi_{k|a_1}}] \\ &= E_1[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{12}} \frac{y_k}{\pi_k}] \\ &= E_1[\sum_{k=1}^{n} \frac{y_k}{\pi_k}], \end{aligned}$$

where $E_2(b_k) = \pi_{k|a_1}$ and $n = n_{11} + n_{12}$.

Let $a_k = \begin{cases} 1 & \text{if } y_k \in n \\ 0 & \text{otherwise} \end{cases}$ be a *Bernoulli* random variable, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned} &= E_1[\sum_{k=1}^{N} \frac{a_k y_k}{\pi_k}] \\ &= \sum_{k=1}^{N} y_k \\ &= \tau, \end{aligned}$$

where $E_1(a_k) = \pi_k$.

To prove equation 3.18 note that the variance of $\hat{\tau}_{\pi ps}^{srs}$ in the initial sample phase is

$$\begin{aligned} V_1E_2E_3(\hat{\tau}_{\pi ps}^{srs}) &= V_1[\sum_{k=1}^{n} \frac{y_k}{\pi_k}] \\ &= V_1[\sum_{k=1}^{N} \frac{a_k y_k}{\pi_k}] \\ &= \sum_{k=1}^{N} \frac{y_k^2 V_1(a_k)}{\pi_k^2} + \sum_{k=1}^{N}\sum_{k\neq l}^{N} (\frac{y_k y_l}{\pi_k \pi_l}) cov(a_k, a_l) \\ &= \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N}\sum_{k\neq l}^{N} (\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}) y_k y_l \\ &= \sum_{k=1}^{N}\sum_{l=1}^{N} (\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}) y_k y_l, \end{aligned}$$

$$\begin{aligned} E_1V_2E_3(\hat{\tau}_{\pi ps}^{srs}) &= E_1V_2[\sum_{k=1}^{n_{11}} \frac{y_k}{\pi_k} + \sum_{k=1}^{n_{12}'} \frac{y_k}{\pi_k \pi_{k|a_1}}] \\ &= E_1V_2[\sum_{k=1}^{n_{12}'} \frac{y_k}{\pi_k^{(1)}}] \\ &= E_1V_2[\sum_{k=1}^{n_{12}} \frac{b_k y_k}{\pi_k^{(1)}}] \\ &= E_1[\sum_{k=1}^{n_{12}} (\frac{y_k}{\pi_k^{(1)}})^2 V_2(b_k) + \sum_{k=1}^{n_{12}}\sum_{k\neq l}^{n_{12}} (\frac{y_k y_l}{\pi_k^{(1)}\pi_l^{(1)}}) cov(b_k, b_l)] \\ &= E_1[\sum_{k=1}^{n_{12}} \frac{y_k^2}{\pi_k^2 \pi_{k|a_1}^2} \pi_{k|a_1}(1-\pi_{k|a_1}) + \sum_{k=1}^{n_{12}}\sum_{k\neq l}^{n_{12}} \frac{(\pi_{kl|a_1} - \pi_{k|a_1}\pi_{l|a_1})}{\pi_k \pi_{k|a_1} \pi_l \pi_{l|a_1}} y_k y_l] \\ &= E_1[\sum_{k=1}^{n_{12}}\sum_{l=1}^{n_{12}} \frac{(\pi_{kl|a_1} - \pi_{k|a_1}\pi_{l|a_1})}{\pi_{k|a_1}\pi_{l|a_1}} \frac{y_k y_l}{\pi_k \pi_l}], \end{aligned}$$

and

$$\begin{aligned} E_1E_2V_3(\hat{\tau}^{srs}_{\pi ps}) &= E_1E_2V_3[\sum_{k=1}^{n_{11}}\frac{y_k}{\pi_k}+\sum_{k=1}^{n_{12}}\frac{y_k}{\pi_k\pi_{k|a_1}}+\sum_{k=1}^{n'_{22}}\frac{y_k}{\pi_k\pi_{k|a_1}\pi_{k|a_2}}] \\ &= E_1E_2V_3[\sum_{k=1}^{n_{22}}\frac{c_ky_k}{\pi^{(2)}_k}] \\ &= E_1E_2[\sum_{k=1}^{n_{22}}\frac{y_k^2\pi_{k|a_2}(1-\pi_{k|a_2})}{\pi_k^2\pi^2_{k|a_1}\pi^2_{k|a_2}}+\sum_{k=1}^{n_{22}}\sum_{k\neq l}^{n_{22}}(\frac{\pi_{kl|a_2}-\pi_{k|a_2}\pi_{l|a_2}}{\pi_k\pi_{k|a_1}\pi_{k|a_2}\pi_l\pi_{l|a_1}\pi_{l|a_2}})y_ky_l] \\ &= E_1E_2\sum_{k=1}^{n_{22}}\sum_{l=1}^{n_{22}}(\pi_{kl|a_2}-\pi_{k|a_2}\pi_{l|a_2})\frac{y_ky_l}{\pi^{(2)}_k\pi^{(2)}_l}, \end{aligned}$$

where $\pi^{(2)}_k = \pi_k\pi_{k|a_1}\pi_{k|a_2}$.

Equation 3.18 follows by lemma 3.2. To prove the unbiased variance estimator, a *Bernoulli* random variable will be used to show the unbiased estimates in each component of this estimator. For the first component, let $a_{kl} = \begin{cases} 1 & \text{if } y_k, y_l \in S' \\ 0 & \text{otherwise} \end{cases}$ be a *Bernoulli* random variable with inclusion probability π^*_{kl} and

$$\sum_{k=1}^{S'}\sum_{l=1}^{S'}\frac{\triangle_{kl}}{\pi^*_{kl}}\tilde{y}_k\tilde{y}_l = \sum_{k=1}^{S'}\sum_{l=1}^{S'}(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l})\frac{y_ky_l}{\pi^*_{kl}} = \sum_{k=1}^{N}\sum_{l=1}^{N}(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l\pi^*_{kl}})a_{kl}y_ky_l,$$

then

$$E[\sum_{k=1}^{S'}\sum_{l=1}^{S'}(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l})\frac{y_ky_l}{\pi^*_{kl}}] = \sum_{k=1}^{N}\sum_{l=1}^{N}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}y_ky_l,$$

since $E(a_{kl}) = \pi^*_{kl}$. Similar, the two last components can be proved for unbiased estimator. Then an unbiased estimator of the population total is given in equation 3.19.

■

In sampling with replacement, a simple random sample of size n is drawn according to the design with probabilities p_k. As with the previous case for sampling without replacement, despite efforts to obtain response y_k from all elements in n, some nonresponse occurs. Sampling units can be classified into two categories. The first category is a response group with size n_{11} while the other is a nonresponse

group with size n_{12}. First nonrespondent subsampling with size $n_{12}^{'} = \frac{n_{12}}{k_1}, k_1 > 0$ is predetermined, is subsampled with probabilities denoted $p_{k|a_1}$. There are only n_{21} responses in this phase. Second subsampling of size $n_{22}^{'} = \frac{n_{22}}{k_2}$ units with probabilities denoted $p_{k|a_2}$, $k_2 > 0$ is predetermined and $n_{22} = n_{12}^{'} - n_{21}$, will be then conducted and response is assumed completely.

Theorem 3.10 *Random Unequal Probability Sampling with Replacement*

In a simple random sample of size n with nonrespondent subsampling, an unbiased estimator of τ is

$$\hat{\tau}_{pps}^{srs} = \frac{1}{n}[\sum_{k=1}^{n_{11}} \frac{y_k}{p_k} + \frac{1}{n_{12}^{'}}(\sum_{k=1}^{n_{21}} \frac{y_k}{p_k p_{k|a_1}} + \frac{1}{n_{22}^{'}} \sum_{k=1}^{n_{22}^{'}}(\frac{y_k}{p_k p_{k|a_1} p_{k|a_2}}))], \tag{3.20}$$

with a variance of

$$V(\hat{\tau}_{pps}^{srs}) = \frac{1}{n}[\sum_{k=1}^{N} \frac{y_k^2}{p_k} - \tau^2] +$$

$$E_1 \frac{1}{n^2 n_{12}^{'}}[\sum_{k=1}^{n_{12}} \frac{y_k^2}{p_k^{(1)}} - \tau^{(1)2}] + E_1 E_2 \frac{1}{n^2 n_{12}^{'2} n_{22}^{'}}[\sum_{k=1}^{n_{22}} \frac{y_k^2}{p_k^{(2)}} - \tau^{(2)2}]. \tag{3.21}$$

An unbiased estimator for the variance of the sample total is

$$v(\hat{\tau}_{pps}^{srs}) = \frac{1}{S^{'}(S^{'}-1)}[\sum_{k=1}^{S^{'}} \frac{y_k^2}{p_k^2} - S^{'}\hat{\tau}^2] +$$

$$\frac{1}{n^2 n_{12}^{'}(n_{12}^{'}-1)}[\sum_{k=1}^{n_{12}^{'}} \frac{y_k^2}{p_k^{(1)2}} - n_{12}^{'}\hat{\tau}^{(1)2}] + \frac{1}{n^2 n_{12}^{'2} n_{22}^{'}(n_{22}^{'}-1)}[\sum_{k=1}^{n_{22}^{'}} \frac{y_k^2}{p_k^{(2)2}} - n_{22}^{'}\hat{\tau}^{(2)2}], \tag{3.22}$$

where $\hat{\tau}^{(1)}$ and $\hat{\tau}^{(2)}$ are the total population estimate in the first and second subsampling respectively, $p_k^{(1)} = p_k p_{k|a_1}$ and $p_k^{(2)} = p_k p_{k|a_1} p_{k|a_2}$, and $S^{'} = n_{11} + n_{21} + n_{22}^{'}$.

Proof: $\hat{\tau}_{pps}^{srs}$ is unbiased estimator if $E(\hat{\tau}) = \tau$. It is noted that

$$E(\hat{\tau}_{pps}^{srs}) = E_1 E_2 E_3 \frac{1}{n}[\sum_{k=1}^{n_{11}} \frac{y_k}{p_k} + \frac{1}{n_{12}^{'}}(\sum_{k=1}^{n_{21}} \frac{y_k}{p_k p_{k|a_1}} + \frac{1}{n_{22}^{'}} \sum_{k=1}^{n_{22}^{'}} \frac{y_k}{p_k p_{k|a_1} p_{k|a_2}})],$$

let c_k be a Binomial random variable with probability $p_{k|a_2}$, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned}
&= E_1E_2E_3\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}(\sum_{k=1}^{n_{21}}\frac{y_k}{p_kp_{k|a_1}}+\frac{1}{n_{22}^{'}}\sum_{k=1}^{n_{22}}\frac{c_ky_k}{p_kp_{k|a_1}p_{k|a_2}})]\\
&= E_1E_2\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}(\sum_{k=1}^{n_{21}}\frac{y_k}{p_kp_{k|a_1}}+\sum_{k=1}^{n_{22}}\frac{y_k}{p_kp_{k|a_1}})]\\
&= E_1E_2\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}\sum_{k=1}^{n_{12}^{'}}\frac{y_k}{p_kp_{k|a_1}}],
\end{aligned}$$

where $E_3(c_k)=n_{22}^{'}p_{k|a_2}$ and $n_{12}^{'}=n_{21}+n_{22}$.

Let b_k be a Binomial random variable with probability $p_{k|a_1}$, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned}
&= E_1E_2\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}\sum_{k=1}^{n_{12}}\frac{b_ky_k}{p_kp_{k|a_1}}]\\
&= E_1E_2\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\sum_{k=1}^{n_{12}}\frac{y_k}{p_k}]\\
&= E_1\frac{1}{n}[\sum_{k=1}^{n}\frac{y_k}{p_k}],
\end{aligned}$$

where $E_2(b_k)=n_{12}^{'}p_{k|a_1}$ and $n=n_{11}+n_{12}$.

Let a_k be a Binomial random variable with probability p_k, then $E(\hat{\tau})$ can be rewritten as

$$\begin{aligned}
&= E_1\frac{1}{n}[\sum_{k=1}^{N}\frac{a_ky_k}{p_k}]\\
&= \sum_{k=1}^{N}y_k\\
&= \tau,
\end{aligned}$$

where $E_1(a_k)=np_k$. To prove equation 3.21 note that

$$\begin{aligned}
V_1E_2E_3(\hat{\tau}_{pps}^{srs}) &= V_1[\frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{p_k}]\\
&= \frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k}-\tau^2],
\end{aligned}$$

$$\begin{aligned} E_1V_2E_3(\hat{\tau}_{pps}^{srs}) &= E_1V_2\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}\sum_{k=1}^{n_{12}^{'}}\frac{y_k}{p_kp_{k|a_1}}] \\ &= E_1\frac{1}{n^2}V_2[\frac{1}{n_{12}^{'}}\sum_{k=1}^{n_{12}^{'}}\frac{y_k}{p_k^{(1)}}] \\ &= E_1\frac{1}{n^2n_{12}^{'}}[\sum_{k=1}^{n_{12}}\frac{y_k^2}{p_k^{(1)}}-\tau^{(1)2}], \end{aligned}$$

where $p_k^{(1)}=p_kp_{k|a_1}$ and $\tau^{(1)}$ is the total population in the first subsampling and

$$\begin{aligned} E_1E_2V_3(\hat{\tau}_{pps}^{srs}) &= E_1E_2V_3\frac{1}{n}[\sum_{k=1}^{n_{11}}\frac{y_k}{p_k}+\frac{1}{n_{12}^{'}}(\sum_{k=1}^{n_{21}}\frac{y_k}{p_k^{(1)}}+\frac{1}{n_{22}^{'}}\sum_{k=1}^{n_{22}^{'}}\frac{y_k}{p_k^{(2)}})] \\ &= E_1E_2\frac{1}{n^2n_{12}^{'2}}V_3[\frac{1}{n_{22}^{'}}\sum_{k=1}^{n_{22}^{'}}\frac{y_k}{p_k^{(2)}}] \\ &= E_1E_2\frac{1}{n^2n_{12}^{'2}n_{22}^{'}}[\sum_{k=1}^{n_{22}}\frac{y_k^2}{p_k^{(2)}}-\tau^{(2)2}], \end{aligned}$$

where $\tau^{(2)}$ is the total population in the second subsampling.

By lemma 3.2 the variance of total estimator is equation 3.21.

Since, $\frac{1}{S^{'}(S^{'}-1)}[\sum_{k=1}^{S^{'}}\frac{y_k^2}{p_k^2}-S^{'}\hat{\tau}_{pps}^2]$ is an unbiased estimate of $\frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k}-\tau^2]$. By theorem 2.7, $\frac{1}{n_{12}^{'}(n_{12}^{'}-1)}[\sum_{k=1}^{n_{12}^{'}}\frac{y_k^2}{p_k^{(1)2}}-n_{12}^{'}\hat{\tau}^{(1)2}]$ and $\frac{1}{n_{22}^{'}(n_{22}^{'}-1)}[\sum_{k=1}^{n_{22}^{'}}\frac{y_k^2}{p_k^{(2)2}}-n_{22}^{'}\hat{\tau}^{(2)2}]$ are an unbiased estimator of $\frac{1}{n_{12}^{'}}[\sum_{k=1}^{n_{12}}\frac{y_k^2}{p_k^{(1)}}-\tau^{(1)2}]$ and $\frac{1}{n_{22}^{'}}[\sum_{k=1}^{n_{22}}\frac{y_k^2}{p_k^{(2)}}-\tau^{(2)2}]$ respectively.

Then unbiased estimator of the total estimate is given in equation 3.22, where $S^{'}=n_{11}+n_{21}+n_{22}^{'}$.

■

In sampling without replacement, a stratified random sample with varying probabilities of selection can be considered a survey plan in which a population is divided into H mutually exclusive and exhaustive strata and a simple random sample of n_h elements is taken within each stratum h with probability $\pi_{hk}, k=1,...,n_h$. Unfortunately, it is rarely possible to achieve total success in obtaining complete data from all of the unit selected. For n_{h11} units the sample survey has obtained full or some information, and for the other n_{h12} units, the survey has obtained no information

at all. A first subsample of size $n'_{h12} = \frac{n_{h12}}{k_{h1}}$ with a positive inclusion probability $\pi_{hk|a_1}$ and $\pi_{hkl|a_1}$, where $k_{h1} > 0$ is predetermined, is conducted but there are only n_{h21} responses. A second subsample of size $n'_{h22} = \frac{n_{h22}}{k_{h2}}$ with a positive inclusion probability $\pi_{hk|a_2}$ and $\pi_{hkl|a_2}$, with $k_{h2} > 0$ fixed in advance and $n_{h22} = n'_{h12} - n_{h21}$, is conducted and the full response is required in this phase.

Theorem 3.11 *Stratified Unequal Probability Sampling without Replacement*

In a stratified random sample of size n with nonrespondent subsampling, an unbiased estimator of τ is

$$\hat{\tau}^{st}_{\pi ps} = \sum_{h=1}^{H} \sum_{k=1}^{S'_h} \frac{y_{hk}}{\pi^*_{hk}}, \tag{3.23}$$

where $S'_h = n_{h11} + n_{h21} + n'_{h22}$ and

$$\pi^*_{hk} = \begin{cases} \pi_{hk} & \text{if } k \in n_{h11}, \\ \pi_{hk}\pi_{hk|a_1} & \text{if } k \in n_{h21}, \\ \pi_{hk}\pi_{hk|a_1}\pi_{hk|a_2} & \text{if } k \in n'_{h22}, \end{cases}$$

with a variance of

$$V(\hat{\tau}^{st}_{\pi ps}) = \sum_{h=1}^{H} \sum_{k=1}^{N_h} \sum_{l=1}^{N_h} \left(\frac{\pi_{hkl} - \pi_{hk}\pi_{hl}}{\pi_{hk}\pi_{hl}}\right) y_{hk} y_{hl} +$$

$$\sum_{h=1}^{H} \sum_{k=1}^{n_{h12}} \sum_{l=1}^{n_{h12}} (\pi_{hkl|a_1} - \pi_{hk|a_1}\pi_{hl|a_1}) \frac{y_{hk} y_{hl}}{\pi^{(1)}_{hk}\pi^{(1)}_{hl}} + \sum_{h=1}^{H} \sum_{k=1}^{n_{h22}} \sum_{l=1}^{n_{h22}} (\pi_{hkl|a_2} - \pi_{hk|a_2}\pi_{hl|a_2}) \frac{y_{hk} y_{hl}}{\pi^{(2)}_{hk}\pi^{(2)}_{hl}}, \tag{3.24}$$

where

$$\pi^{(1)}_{hk} = \pi_{hk}\pi_{hk|a_1} \text{ and } \pi^{(2)}_{hk} = \pi_{hk}\pi_{hk|a_1}\pi_{hk|a_2}.$$

An unbiased estimator for the variance of the sample total is

$$v(\hat{\tau}^{st}_{\pi ps}) = \sum_{h=1}^{H} \sum_{k=1}^{S'_h} \sum_{l=1}^{S'_h} \frac{\triangle_{hkl}}{\pi^*_{hkl}} \tilde{y}_{hk} \tilde{y}_{hl} +$$

$$\sum_{h=1}^{H} \sum_{k=1}^{n_{h21}} \sum_{l=1}^{n_{h21}} \frac{\triangle_{hkl|a_1}}{\pi_{hkl|a_1}} \tilde{y}^{(1)}_{hk} \tilde{y}^{(1)}_{hl} + \sum_{h=1}^{H} \sum_{k=1}^{n'_{h22}} \sum_{l=1}^{n'_{h22}} \frac{\triangle_{hkl|a_2}}{\pi_{hkl|a_2}} \tilde{y}^{(2)}_{hk} \tilde{y}^{(2)}_{hl}, \tag{3.25}$$

where

$$\triangle_{hkl} = \pi_{hkl} - \pi_{hk}\pi_{hl}, \triangle_{hkl|a_1} = \pi_{hkl|a_1} - \pi_{hk|a_1}\pi_{hl|a_1}, \triangle_{hkl|a_2} =$$
$$\pi_{hkl|a_2} - \pi_{hk|a_2}\pi_{hl|a_2}, \tilde{y}_{hk} = \frac{y_{hk}}{\pi_{hk}}, \tilde{y}_{hk}^{(1)} = \frac{y_{hk}}{\pi_{hk}^{(1)}}, \tilde{y}_{hk}^{(2)} = \frac{y_{hk}}{\pi_{hk}^{(2)}}$$

and

$$\pi_{hkl}^{*} = \begin{cases} \pi_{hkl} & \text{if } k,l \in n_{h11}, \\ \pi_{hkl}\pi_{hk|a_1} & \text{if } k \in n_{h21}, l \in n_{h11}, \\ \pi_{hkl}\pi_{hkl|a_1} & \text{if } k,l \in n_{h21}, \\ \pi_{hkl}\pi_{hk|a_1}\pi_{hk|a_2} & \text{if } k \in n_{h22}^{'}, l \in n_{h11}, \\ \pi_{hkl}\pi_{hl|a_1}\pi_{hk|a_2} & \text{if } k \in n_{h22}^{'}, l \in n_{h21}, \\ \pi_{hkl}\pi_{hkl|a_1}\pi_{hkl|a_2} & \text{if } k,l \in n_{h22}^{'}, \end{cases}$$

Proof: A stratified random sample is taken in the same way as a simple random sample, but the sampling is done separately and independently within each stratum. By the results in theorem 3.9 and the reason above, an unbiased estimator for a population total and its variance can be done by adding the estimator in each stratum together as

$$\hat{\tau}_{\pi ps}^{st} = \sum_{h=1}^{H} \hat{\tau}_{h,\pi ps}^{srs} \text{ and } V(\hat{\tau}_{\pi ps}^{st}) = \sum_{h=1}^{H} V(\hat{\tau}_{h,\pi ps}^{srs}).$$

■

For a stratified random sample with varying probabilities of selection and sampling with replacement the population is divided into H mutually exclusive and exhaustive strata and a simple random sample of n_h elements is taken within each stratum h with probability $p_{hk}, k = 1, ..., n_h$. Again, it is rarely possible to achieve total success in obtaining complete data from all of the unit selected. For n_{h11} units the sample survey has obtained full or some information, and for other n_{h12} units, the survey has obtained no information at all. A first subsample of size $n_{h12}^{'} = \frac{n_{h12}}{k_{h1}}$ with a probability $p_{hk|a_1}$, with $k_{h1} > 0$ is predetermined, is conducted but there are only n_{h21} responses. A second subsample of size $n_{h22}^{'} = \frac{n_{h22}}{k_{h2}}$ with a probability

$p_{hk|a_2}$, with $k_{h2} > 0$ fixed in advance and $n_{h22} = n'_{h12} - n_{h21}$, is conducted and the full response is required in this phase.

Theorem 3.12 *Stratified Unequal Probability Sampling with Replacement*

In a stratified random sample of size n with nonrespondent subsampling, an unbiased estimator of τ is

$$\hat{\tau}^{st}_{pps} = \sum_{h=1}^{H} \frac{1}{n_h}[\sum_{k=1}^{n_{h11}} \frac{y_{hk}}{p_{hk}} + \frac{1}{n'_{h12}}(\sum_{k=1}^{n_{h21}} \frac{y_{hk}}{p_{hk}p_{hk|a_1}} + \frac{1}{n'_{h22}}(\sum_{k=1}^{n'_{h22}} \frac{y_{hk}}{p_{hk}p_{hk|a_1}p_{hk|a_2}}))], \quad (3.26)$$

with a variance of

$$V(\hat{\tau}^{st}_{pps}) = \sum_{h=1}^{H} \frac{1}{n_h}[\sum_{k=1}^{N_h} \frac{y^2_{hk}}{p_{hk}} - \tau_h^2] +$$

$$\sum_{h=1}^{H} E_1 \frac{1}{n_h^2 n'_{h12}}[\sum_{k=1}^{n_{h12}} \frac{y^2_{hk}}{p^{(1)}_{hk}} - \tau_h^{(1)2}] + \sum_{h=1}^{H} E_1 E_2 \frac{1}{n_h^2 n'^2_{h12} n'_{h22}}[\sum_{k=1}^{n_{h22}} \frac{y^2_{hk}}{p^{(2)}_{hk}} - \tau_h^{(2)2}], \quad (3.27)$$

where $\tau_h^{(1)}$ and $\tau_h^{(2)}$ are the total population in the first and second subsampling of stratum h respectively.

An unbiased estimator for the variance of the sample total is

$$v(\hat{\tau}^{st}_{pps}) = \sum_{h=1}^{H} \frac{1}{S'_h(S'_h - 1)}[\sum_{k=1}^{S'_h} \frac{y^2_{hk}}{p^2_{hk}} - S'_h \hat{\tau}_h^2] +$$

$$\sum_{h=1}^{H} \frac{1}{n_h^2 n'_{h12}(n'_{h12} - 1)}[\sum_{k=1}^{n'_{h12}} \frac{y^2_{hk}}{p^{(1)2}_{hk}} - n'_{h12}\hat{\tau}_h^{(1)2}] + \sum_{h=1}^{H} \frac{1}{n_h^2 n'^2_{h12} n'_{h22}(n'_{h22} - 1)}[\sum_{k=1}^{n'_{h22}} \frac{y^2_{hk}}{p^{(2)2}_{hk}} - n'_{h22}\hat{\tau}_h^{(2)2}], \quad (3.28)$$

where $\hat{\tau}_h^{(1)}$ and $\hat{\tau}_h^{(2)}$ are the total population estimate in the first and second subsampling of stratum h respectively, $p^{(1)}_{hk} = p_{hk}p_{hk|a_1}$ and $p^{(2)}_{hk} = p_{hk}p_{hk|a_1}p_{hk|a_2}$.

Proof: A stratified random sample is taken in the same way as a simple random sample, but the sampling is done separately and independently within each stratum. By the results in theorem 3.10 and the reason above, an unbiased estimator for a population total and its variance can be done by adding the estimator in each stratum together as

$$\hat{\tau}^{st}_{pps} = \sum_{h=1}^{H} \hat{\tau}^{srs}_{h,pps} \text{ and } V(\hat{\tau}^{st}_{pps}) = \sum_{h=1}^{H} V(\hat{\tau}^{srs}_{h,pps}).$$

■

Chapter 4

Weighting Adjustments

Strategies for dealing with unit nonresponse described in the previous chapter are those used at the planning stage or during the period of data collection. Nonrespondence can be dealt with as a preliminary step in data analysis or as part of an evaluation of the study protocol after analysis has been done. Weighting adjustment methods described in this chapter and imputation methods in chapter 5 are those used for data analysis in the process of formulating and producing survey estimates. Section 4.1 gives an overview of weighing adjustment procedures. Section 4.2 presents the notation used in this chapter. Section 4.3 presents some theorems on weighting adjustment procedures.

4.1 Overview

Weighting adjustments are primarily used to compensate for unit nonresponse. The essence of all weighting adjustment procedures is to increase the weights of specified respondents so that they represent the nonrespondents. The procedures require auxiliary information on either the nonrespondents or the total population.

For example, in *SRSWOR* with full response, the sampling weights are the re-

ciprocals of the probabilities of selection, e.g. an estimator of the population total is $\sum_{k=1}^{n} w_k y_k$, where $w_k = \frac{N}{n}$. If nonresponse occurs, then weights might be $\frac{1}{(\pi_k \hat{\phi}_k)}$ as shown below:

Let I_k be an indicator variable equal to one if the unit k is selected, and zero otherwise. The probability that unit I_k is selected, $P(I_k = 1)$ is π_k. If the response indicator random variable R_k is assumed to be independent of I_k then the probability that unit k will be measured is P(unit k selected in sample and responds) $= \pi_k \phi_k$ where $\phi_k = P(R_k = 1)$. The ϕ_k is estimated by $\hat{\phi}_k$ for each unit in the sample, using auxiliary information that is assumed known for all units in the selected sample. The final weight for a respondent is then $\frac{1}{(\pi_k \hat{\phi}_k)}$.

The above weighting method assumes that the response probabilities can be estimated from variables known for all units. These type of methods are said to have a missing at random (*MAR*) mechanism. Most survey methods for compensating nonresponse assume *MAR* mechanism for nonresponse. However, the *MAR* mechanism can often cause bias. *Oh & Scheuren* (1983), *Little* (1986), *Kalton & Kasprzyk* (1986), *Holt & Elliot* (1991) and *Lehtonen & Pahkinen* (1995) propose various methods of weighting which are used to reduce nonresponse bias. The methods may be categorised as those i) using auxiliary population information, ii) using auxiliary information for the intended respondents, and iii) using no auxiliary information. These weighting adjustment methods are based on response distribution modelling.

A response model, which is a set of assumptions about the true unknown response distribution, is a tool used to construct an estimator. Such an estimator will have commendable properties, such as unbiasedness or approximate unbiasedness provided the response model coincides with the actual response distribution, but not necessarily otherwise (*Oh & Scheuren*, 1983). Two response models are commonly used: naive models and response homogeneity group (*RHG*) models. Naive response models can be described as models for data missing at ran-

dom throughout the population in such a way that $P(R_k = 1|S) = \phi_k = \phi$ and $P(R_k = 1, R_i = 1|S) = \phi_k\phi_i = \phi^2$ for all k and $i \in S$ and every sample S and ϕ is a unknown constant ($\phi > 0$). Suppose now that the mean estimator that would have been used in the case of full response is

$$\hat{\mu} = \hat{\mu}_S = \frac{\sum_S y_k/\pi_k}{\sum_S 1/\pi_k},$$

where π_k is the probability that the sampling unit k is selected. If the naive model is adopted, an appropriate modification of $\hat{\mu}$ is

$$\hat{\mu}_m = \frac{\sum_{k=1}^m y_k/\pi_k\phi}{\sum_{k=1}^m 1/\pi_k\phi} = \frac{\sum_{k=1}^m y_k/\pi_k}{\sum_{k=1}^m 1/\pi_k},$$

where m is the response size. The only difference between $\hat{\mu}_m$ and $\hat{\mu}$ is the summation over the response size m instead of over the entire sample S. The nonresponse is simply ignored. Thus for equal probability sampling the mean estimator for the full response and for the response data are

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^n y_k,$$

and

$$\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^m y_k.$$

For a more realistic response model, the realized sample S is partitioned into a contingency table with $H_S \times L_S$ cells, S_{hl}, $h = 1, ..., H_S$ and $l = 1, ..., L_S$. Given S, a response probability is assumed that is the same within the cell S_{hl} but may be different among cells. Elements are assumed to respond independently of each other. Thus the following assumptions are made:

For every S and for $h, h' = 1, ..., H_S$, $l, l' = 1, ..., L_S$. $P(k \in S_{hl}|S) = \phi_{hl} > 0$ for all $k \in S_{hl}$ and $P(k \in S_{hl}, i \in S_{h'l'}|S) = \phi_{hl}\phi_{h'l'}$ for all $k \in S_{hl}$ and $i \in S_{h'l'}$.

This simple assumption leads to a powerful family of response models called random homogeneity group or *RHG* models. *RHG* models are sub-classed into four different types of methods: (i) weighting cell methods or sample-based adjustment methods where the sample proportion is used for weighting factor, (ii) post-stratification methods or population-based adjustment methods where the population proportion is used for weighting factor, (iii) raking ratio methods which the raked number is used for weighting factor (more details discussed below) and (iv) general-based adjustment methods where the response proportion is used for weighting factor.

If *SRSWOR* were used and there were some nonrespondents, then *RHG* models are applied with the four methods above. This would give the following estimators:

i) *Sample-based adjustment* method: Here $\pi_{hlk} = \frac{n}{N}$ and $\hat{\phi}_{hl} = \frac{m_{hl}}{n_{hl}}$ so that

$$\begin{aligned}\hat{\mu}_s &= \frac{\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{m_{hl}} y_{hlk}/(\pi_{hlk}\hat{\phi}_{hl})}{\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} 1/(\pi_{hlk}\hat{\phi}_{hl})} \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hlm},\end{aligned}$$

where $\hat{\mu}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}} y_{hlk}$ is the response mean, m_{hl} and n_{hl} are the response size and sample size respectively in cell *hl*.

ii) *Population-based adjustment* method: If the population proportions $\frac{N_{hl}}{N}$ in each cell are known, an alternative to $\hat{\mu}_s$ is the post-stratified mean,

$$\hat{\mu}_p = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm},$$

where $\frac{n_{hl}}{n}$ in (i) is replaced by $\frac{N_{hl}}{N}$.

iii) *Raking ratio* method: This method can be applied when the population count in cell *hl* is unknown but the marginal counts for X_1 and X_2, where X_1 and X_2 are auxiliary variables corresponding to the two-way classification (contingency) table, $N_{h.} = \sum_{l=1}^{L_S} N_{hl}$ and $N_{.l} = \sum_{h=1}^{H_S} N_{hl}$, are known for all h and l,

from published census data. The N_{hl} can be estimated by using the raking ratio method suggested by *Deming & Stephan* (1940) who used it in census to ensure a complete census data set. *Brackston & Rao* (1976) further developed the theory. *Oh & Scheuren* (1983), *Little & Rubin* (1987), *Deville, Sarndal & Sautory* (1993) also describe raking ratio estimates for nonresponse.

In the raking ratio method *raked cell counts* N_{hl}^* are used to replace N_{hl} in the calculation of $\hat{\mu}_p$. The raked cell counts can be calculated by an *iterative proportional fitting* procedure, where current estimates are scaled by row or column factors to match the marginal total $N_{h.}$ or $N_{.l}$, respectively. *Little & Rubin* (1987) give the following raking ratio procedures. The first step computes

$$(1): \qquad N_{hl}^{(1)} = n_{hl}(N_{h.}/n_{h.}),$$

so that $N_{h.}^{(1)} = N_{h.}$. Then

$$(2): \qquad N_{hl}^{(2)} = N_{hl}^{(1)}(N_{.l}/N_{.l}^{(1)})$$

is computed so that $N_{.l}^{(2)} = N_{.l}$. Then

$$(3): \qquad N_{hl}^{(3)} = N_{hl}^{(2)}(N_{h.}/N_{h.}^{(2)})$$

is computed so that $N_{h.}^{(3)} = N_{h.}$. The procedures in steps (2) and (3) are repeated until N_{hl}^i convergence. Convergence and statistical properties of this procedure are discussed by *Ireland & Kullback* (1968), who show, in particular, that the raking ratio estimates N_{hl}^*/N of the cell proportions are optimal asymptotically normal estimates under a multinomial assumption for the cell counts n_{hl}, and as such are asymptotically equivalent to the maximum likelihood estimate under the multinomial model. The raking ratio estimator of μ is

$$\hat{\mu}_r = \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{N_{hl}^*}{N} \hat{\mu}_{hlm},$$

which might be expected to have variance properties somewhere between $\hat{\mu}_p$ and $\hat{\mu}_s$. If the sample sizes in each cell are large enough, the raking ratio estimator is approximately unbiased. However, this estimator is not defined for any cell hl when $m_{hl} = 0$ and $n_{hl} \neq 0$. In this situation some other estimator of the mean for that cell is required (*Little & Rubin, 1987*).

iv) *General-based adjustment* method: Sometimes both the population cell counts and the marginal population cell counts are unknown. An alternative estimator is the general-based adjustment method proposed by *Lehtonen & Pahkinen* (1995). They suggest using sample response proportions $\frac{m_{hl}}{m}$ as an alternative for *RHG* models giving

$$\hat{\mu}_g = \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{m_{hl}}{m} \hat{\mu}_{hlm}.$$

With each of these models it is possible to use various sampling design. In this thesis, naive models for weighting adjustments are considered in conjunction with simple random sampling, stratified random sampling and post-stratified random sampling designs. *RHG* models are considered in conjunction with simple random sampling only since stratified and post-stratified random sampling design need more auxiliary information for sub-classification in each cell and this adds to the complication. The bias-removal method is one I propose for *RHG* models under unequal probability sampling with replacement to compensate for bias in the usual estimates (see section 4.3.2 for more details).

4.2 Notation

The following notations are used in this chapter:

1. $\hat{\mu}_m^d$ is an estimated mean of a characteristic of interest of the sampling design "d" and model "m". The design "d" can be "srs", "st" or "pt" for simple random sampling, stratified random sampling and post-stratified random sampling respectively. The model "m" can be "na", "s", "p", "r", "g" or "u" for naive model, sample-based adjustment, population-based adjustment, raking ratio adjustment, general-based adjustment and bias-removal respectively.

2. $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(y_k-\mu)^2$ and $S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k-\mu)^2$ are the population variance of the study variable Y for sampling with and without replacement respectively.

3. $s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k-\hat{\mu}_m^d)^2$ is the sample variance of the study variable Y under design "d" and model "m".

4. $\hat{\tau}_m^{d,sel}$ is a total estimate of a study variable Y for design "d" and model "m" in the sampling plan where "d" and "m" are the same symbols as in (1) and "sel" can be πps or pps for unequal probability sampling without and with replacement respectively.

5. S and S_{hl} are the set Y_k drawn in a sample and in a sample of stratum hl respectively, usually n and n_{hl} of them. We assume $n_{hl} < n$ and $\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} n_{hl} = n$.

6. R and R_{hl} are the set of respondents in a sample and in a sample of stratum hl respectively, usually m and m_{hl} of them. We assume $m_{hl} < m$ and $\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} m_{hl} = m$.

4.3 Weighting Adjustment Procedure Theory

In this section, theorems about mean estimators are proved for equal probability sampling while total estimators are proved for unequal probability sampling. These theorems parallel the basic sampling theorems in section 2.3.3. All theorems in simple random sampling are assumed to use naive and *RHG* models but stratified random sampling and post-stratified random sampling use only the naive model. Notations used in these theorems are given in section 4.2.

In all these theorems we estimate the mean $\mu = \frac{1}{N}\sum_{k=1}^{n} Y_k$, or the total $\tau =$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{N} I_k Y_k,$$

where I_k is the k^{th} sample indicator. For the mean estimator with nonresponse this can be written as

$$\hat{\mu} = \frac{1}{m}\sum_{k=1}^{N} I_k R_k Y_k,$$

where R_k is the k^{th} response indicator. In all cases if $\hat{\theta}$ is denoted to be the estimator $\hat{\mu}$, $\hat{\tau}$, $\tilde{\mu}$ or $\tilde{\tau}$, we see that θ is a function of $\mathbf{Y}$, $\mathbf{I}$ and $\mathbf{R}$. In every cases $\mathbf{I}$ is dependent of the auxiliary variable $\mathbf{X}$, i.e., unequal probability sampling. But in all cases the distribution of the estimator combined here is that of $\mathbf{I}$ and $\mathbf{R}$ given $\mathbf{Y}$, $\mathbf{X}$ and n and sometimes m. In stratified and post-stratified random sampling, n and m may be a vector.

Section 2.6.4.1 presents quasi-randomisation approach for inference with nonresponse. Lemma 4.1 shows the procedure of expectation and variance of the estimator in this inference approach.

Before the theorems for weighting adjustment methods are introduced, the following six lemmas are proved as these are used in the proof of theorems 4.1-4.27. In these lemmas and following theorems,

$f_I(.)$ is the marginal distribution of $\mathbf{I}$ given $\mathbf{Y}, \mathbf{X}, n$.

$f_R(.)$ is the marginal distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, n, m$.

$f_{IR}(.)$ is the joint distribution of $\mathbf{I}$ and $\mathbf{R}$ given $\mathbf{Y}, \mathbf{X}, n, m$.

$E_I(\hat{\theta})$ is the expectation of $\hat{\theta}$ with respect to the distribution of I given $\mathbf{Y}, \mathbf{X}, n$.

$E_R(\hat{\theta})$ is the expectation of $\hat{\theta}$ with respect to the distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, n, m$.

$V_I(\hat{\theta})$ is the variance of $\hat{\theta}$ with respect to the distribution of $\mathbf{I}$ given $\mathbf{Y}, \mathbf{X}, n$.

$V_R(\hat{\theta})$ is the variance of $\hat{\theta}$ with respect to the distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, n, m$.

$\mathbf{n}, \mathbf{m}$ is a vector of sample size and response size in post-stratified random sample with $\mathbf{n} = (n_1, ..., n_H)$ and $\mathbf{m} = (m_1, ..., m_H)$ respectively.

$f_I(.|\mathbf{n}, \mathbf{m})$ is the marginal distribution of $\mathbf{I}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ in post-stratum h for $h = 1, ..., H$.

$f_R(.|\mathbf{n}, \mathbf{m})$ is the marginal distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}$ in post-stratum h for $h = 1, ..., H$.

$f_{IR}(.|\mathbf{n}, \mathbf{m})$ is the joint distribution of $\mathbf{I}$ and $\mathbf{R}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}$ in post-stratum h for $h = 1, ..., H$.

$E_I(\hat{\theta}|\mathbf{n}, \mathbf{m})$ is the expectation of $\hat{\theta}$ with respect to the distribution of $\mathbf{I}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ in post-stratum h for $h = 1, ..., H$.

$E_R(\hat{\theta}|\mathbf{n}, \mathbf{m})$ is the expectation of $\hat{\theta}$ with respect to the distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}$ in post-stratum h for $h = 1, ..., H$.

$V_I(\hat{\theta}|\mathbf{n}, \mathbf{m})$ is the variance of $\hat{\theta}$ with respect to the distribution of $\mathbf{I}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ in post-stratum h for $h = 1, ..., H$.

$V_R(\hat{\theta}|\mathbf{n}, \mathbf{m})$ is the variance of $\hat{\theta}$ with respect to the distribution of $\mathbf{R}$ given $\mathbf{I}, \mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}$ in post-stratum h for $h = 1, ..., H$.

Lemma 4.1 *Expectation and Variance in Quasi-randomisation*

The expectation of an unbiased estimator $\hat{\theta}$ given $\mathbf{Y}, \mathbf{X}, n$ and m in quasi-randomisation theory is

$$E(\hat{\theta}) = E(\hat{\theta}|\mathbf{Y}, \mathbf{X}, n, m) = E_I E_R(\hat{\theta}).$$

The variance of an estimator $\hat{\theta}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ and $\mathbf{m}$ in quasi-randomisation theory is

$$V(\hat{\theta}) = V(\hat{\theta}|\mathbf{Y}, \mathbf{X}, n, m) = V_I E_R(\hat{\theta}) + E_I V_R(\hat{\theta}).$$

Proof: Let $\hat{\theta}$ be a function of the variable $(\mathbf{Y}, n, m)$. $\hat{\theta}$ can be written as a function of the sample indicator $\mathbf{I}$, the response indicator $\mathbf{R}$, variable $\mathbf{Y}$, n and m, $\hat{\theta}(\mathbf{I}, \mathbf{R}, \mathbf{Y}, n, m)$.

Now Y_k is a fixed value in the population. $\mathbf{I}$ and $\mathbf{R}$ are assumed conditionally independent. The expectation of an unbiased estimator $\hat{\theta}$ is then

$$\begin{aligned} E(\hat{\theta}) &= \sum_I \sum_R \hat{\theta} f_{(I,R)}(i,r) \\ &= \sum_I f_I(i) \sum_R \hat{\theta} f_R(r|i) \\ &= \sum_I f_I(i) E_R(\hat{\theta}|I) \\ &= E_I E_R(\hat{\theta}). \end{aligned}$$

The variance of $\hat{\theta}$ is

$$\begin{aligned} V(\hat{\theta}) &= E_I E_R(\hat{\theta} - \theta)^2 \\ &= E_I E_R[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2] \\ &= E_I[E_R\hat{\theta}^2 - 2\theta E_R\hat{\theta} + \theta^2] \\ &= E_I[\{E_R(\hat{\theta})\}^2 + V_R(\hat{\theta}) - 2\theta E_R(\hat{\theta}) + \theta^2] \\ &= E_I[E_R(\hat{\theta}) - \theta]^2 + E_I V_R(\hat{\theta}) \\ &= V_I E_R(\hat{\theta}) + E_I V_R(\hat{\theta}). \end{aligned}$$

More details in general are in *Mood, Graybill* and *Boes* (1983).

■

Lemma 4.2 *Expectation for Response Units*

In sampling with or without replacement for a sample of size n with m respondents, $E_R(\hat{\mu}_m) = \hat{\mu}$, where $\hat{\mu}_m = \frac{1}{m}\sum_{k\in R} y_k$ and $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} y_k$.

Proof: Let R_k, $k = 1, ..., n$, be the response indicator when $\mathbf{I}$, n and m are given. Assume that exchangeability, i.e. MAR, is used within selected sample. In this case y_k, n and m are determined and

$$f_R(1|\mathbf{I}, n, m) = p(R_k = 1|\mathbf{I}, n, m) = \frac{m}{n}.$$

Thus,

$$E_R(\hat{\mu}_m) = E_R \frac{1}{m} \sum_{k\in R} y_k$$

$$\begin{aligned} &= E_R \frac{1}{m} \sum_{k=1}^{n} R_k y_k \\ &= \frac{1}{m} \sum_{k=1}^{n} \frac{m}{n} y_k \\ &= \frac{1}{n} \sum_{k=1}^{n} y_k \\ &= \hat{\mu}. \end{aligned}$$

■

Lemma 4.3 *Expectation for Sample Units*

In sampling with or without replacement for a sample of size n, $E_I(\hat{\mu}) = \mu$, where $\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} y_k$ and $\mu = \frac{1}{N}\sum_{k=1}^{N} Y_k$.

Proof: i) *Sampling with replacement* case: Let t_k be the number of times Y_k appears in the sample. $P[t_k = i] \sim B(n, 1/N)$ and $E_I(t_k) = \frac{n}{N}$. Then,

$$\begin{aligned} E_I(\hat{\mu}) &= E_I \frac{1}{n} \sum_{k=1}^{n} y_k \\ &= E_I \frac{1}{n} \sum_{k=1}^{N} t_k Y_k \\ &= \frac{1}{n} \sum_{k=1}^{N} \frac{n}{N} Y_k \\ &= \mu. \end{aligned}$$

ii) *Sampling without replacement* case: Let I_k, $k = 1, ..., N$, be the sample indicator with $I_k = 1$ if Y_k is chosen in a sample, then

$$\begin{aligned} E_I(\hat{\mu}) &= E_I \frac{1}{n} \sum_{k=1}^{n} y_k \\ &= E_I \frac{1}{n} \sum_{k=1}^{N} I_k Y_k \\ &= \frac{1}{n} \sum_{k=1}^{N} \frac{n}{N} Y_k \\ &= \mu, \end{aligned}$$

since $E_I(I_k) = \frac{n}{N}$.

See for example *Haslett* (1985) for this proof in a more general context.

■

Lemma 4.4 *Expectation and Variance for Post-stratified Sample in Quasi-randomisation*

The expectation of an unbiased estimator $\hat{\theta}$ for post-stratified sample in quasi-randomisation theory is, for fixed $\mathbf{n}$ and $\mathbf{m}$,

$$E(\hat{\theta}|\mathbf{n},\mathbf{m}) = E(\hat{\theta}|Y,X,\mathbf{n},\mathbf{m}) = E_I E_R(\hat{\theta}|\mathbf{n},\mathbf{m}).$$

The variance of an estimator $\hat{\theta}$ for post-stratified sample in quasi-randomisation theory is

$$V(\hat{\theta}|\mathbf{n},\mathbf{m}) = V(\hat{\theta}|Y,X,\mathbf{n},\mathbf{m}) = V_I E_R(\hat{\theta}|\mathbf{n},\mathbf{m}) + E_I V_R(\hat{\theta}|\mathbf{n},\mathbf{m}).$$

Proof: Proof for the expectation and variance of the estimator for post-stratified random sampling is similar to that for lemma 4.1. ■

Lemma 4.5 *Expectation for Response Units in Post-stratum h*

In sampling with or without replacement for a sample of size n with m respondents, n_h and m_h are classified in post-stratum h, $E_R(\hat{\mu}_{hm}|\mathbf{n},\mathbf{m}) = \hat{\mu}_h$, where $\hat{\mu}_{hm} = \frac{1}{m_h}\sum_{k\in R_h} y_{hk}$ and $\hat{\mu}_h = \frac{1}{n_h}\sum_{k=1}^{n_h} y_{hk}$.

Proof: Proof for the expectation of response units in post-stratum h is similar to that for lemma 4.2

■

Lemma 4.6 *Expectation for Sample Units in Post-stratum h*

In sampling with or without replacement for a sample of size n, n_h are classified in post-stratum h, $E_I(\hat{\mu}_h|\mathbf{n},\mathbf{m}) = \mu_h$, where $\hat{\mu}_h = \frac{1}{n_h}\sum_{k=1}^{n_h} y_{hk}$ and $\mu_h = \frac{1}{N_h}\sum_{k=1}^{N_h} Y_{hk}$.

Proof: Proof for the expectation of sample units in post-stratum h is similar to that for lemma 4.3.

■

Theorems for the naive model and the *RHG* models are presented separately in section 4.3.1 and 4.3.2 respectively.

To obtain the estimated variance for equal probability sampling design in theorems 4.1-4.2 replace σ^2 or S^2 with $s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k - \hat{\mu}_m)^2$, where $\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} y_k$ and m is the response size in the sample. The estimated variance in theorem 4.3-4.6 can be replaced σ_h^2 or S_h^2 with $s_{hm}^2 = \frac{1}{m_h-1}\sum_{k=1}^{m}(y_{hk} - \hat{\mu}_{hm})^2$, where $\hat{\mu}_{hm} = \frac{1}{m_h}\sum_{k=1}^{m_h} y_{hk}$ and m_h is the response size in the sample of stratum h. The estimated variances are similarly found in theorems 4.11,4.12,4.15,4,16,4.19,4.20,4.23 and 4.24. For remaining theorems, estimates of variance are made as in theorem 2.7 or 2.8.

Note that theorems (below) concerning with equal probability sampling with replacement, when weighting adjustment procedure is applied, are given the variance of mean estimator as high as approximately two times of the variance when there are no nonrespondents, $V(\hat{\mu}_m) = \frac{2\sigma^2}{n}$. This could be one of the reasons that sampling with replacement is not useful to conduct the survey and, in this case, ignored nonrespondents is better than using weighting adjustment method when sampling with replacement is applied, $V(\hat{\mu}_m) = \frac{\sigma^2}{m}$, where $m < n$.

4.3.1 Weighting Adjustment Methods with the Naive Model

Simple random sampling is the method of selecting the units from the population in such a way that all possible samples have the same chance of being selected. Usually, units from the population are drawn one by one. If the unit selected at any particular draw is replaced back in the population before the next units is drawn, the procedure is called *with replacement (WR) sampling.* However, if the sampling is done in such a way that each population units can only be selected once in a sample, the procedure is called *simple random sampling without replacement* or *SRSWOR.*

Under the naive model, *Oh & Scheuren* (1983) give the result of theorem 4.1 without proof. I supply a proof and also in theorems 4.2-4.10. I extend these results for weighting adjustment procedure under the naive model using the basic

survey designs described in section 2.3.3. A sample of size n is assumed there are m respondents.

Theorem 4.1 *Simple Random Sampling without Replacement under the naive model*

In Simple random sampling without replacement under the naive model, an unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{srs} = \frac{1}{m}\sum_{k\in R} y_k, \tag{4.1}$$

with a variance of

$$V(\hat{\mu}_{na}^{srs}) = (\frac{1}{m} - \frac{1}{N})S^2. \tag{4.2}$$

Proof: By lemmas 4.2 and 4.3,

$$\begin{aligned} E(\hat{\mu}_{na}^{srs}) &= E_I E_R[\frac{1}{m}\sum_{k\in R} y_k] \\ &= E_I[E_R(\hat{\mu}_m)] \\ &= E_I(\hat{\mu}) \\ &= \mu. \end{aligned}$$

To prove equation 4.2, note that by lemmas 4.2, 4.3 and theorem 2.2

$$\begin{aligned} V_I E_R(\hat{\mu}_{na}^{srs}) &= V_I E_R[\hat{\mu}_m] \\ &= V_I[\hat{\mu}] \\ &= \frac{N-n}{nN}S^2 \\ &= (\frac{1}{n} - \frac{1}{N})S^2, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_{na}^{srs}) &= E_I V_R[\hat{\mu}_m] \\ &= E_I[\frac{n-m}{mn}]s^2 \\ &= (\frac{1}{m} - \frac{1}{n})S^2. \end{aligned}$$

Equation 4.2 follows by lemma 4.1.

■

Theorem 4.2 *Simple Random Sampling with Replacement under the naive model*

In simple random sampling with replacement under the naive model, an unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{srs} = \frac{1}{m}\sum_{k\in R} y_k, \tag{4.3}$$

with a variance of

$$V(\hat{\mu}_{na}^{srs}) = \sigma^2(\frac{1}{n} + \frac{1}{m}). \tag{4.4}$$

Proof: An unbiased estimator of the population mean is proved as in theorem 4.1.

Equation 4.4 also follows the proof of theorem 4.1 with the use theorem 2.1 rather than 2.2 giving $V_I E_R(\hat{\mu}_{na}^{srs}) = \frac{\sigma^2}{n}$ and $E_I V_R(\hat{\mu}_{na}^{srs}) = \frac{\sigma^2}{m}$

■

The procedure of partitioning the population into groups, called strata, and then drawing a sample independently from each stratum, is known as *stratified sampling.* Since the samples from different strata are selected independently, each stratum can be treated as a separate population. For more details and notation used see section 2.3.3. The following two theorems deal with sampling selection with and without replacement in stratified random sampling with a naive model.

Theorem 4.3 *Stratified Random Sampling without Replacement under the naive model*

In stratified random sampling without replacement under the naive model, an unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{st} = \sum_{h=1}^{H}\sum_{l=1}^{L}\frac{N_{hl}}{Nm_{hl}}\sum_{k\in R_{hl}} y_{hlk}, \tag{4.5}$$

with a variance of

$$V(\hat{\mu}_{na}^{st}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}} - \frac{1}{N_{hl}})S_{hl}^2. \tag{4.6}$$

Proof: By theorem 4.1, for each stratum $E(\hat{\mu}_{hlm}) = \frac{1}{N_{hl}} \sum_{k=1}^{N_{hl}} Y_{hlk}$, where $\hat{\mu}_{hlm} = \frac{1}{m_{hl}} \sum_{k \in R_{hl}} y_{hlk}$. Hence

$$\begin{aligned} E(\hat{\mu}_{na}^{st}) &= \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{N_{hl}}{N} E(\hat{\mu}_{hlm}) \\ &= \frac{1}{N} \sum_{h=1}^{H} \sum_{l=1}^{L} \sum_{k=1}^{N_{hl}} Y_{hlk} \\ &= \mu. \end{aligned}$$

Since the selections in different strata are independent and the variance in each stratum is by theorem 4.1, $(\frac{1}{n_{hl}} - \frac{1}{N_{hl}})S_{hl}^2$. Equation 4.6 follows.

■

Theorem 4.4 *Stratified Random Sampling with Replacement under the naive model*

In stratified random sampling with replacement under the naive model, an unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{st} = \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{N_{hl}}{N m_{hl}} \sum_{k \in R_{hl}} y_{hlk}, \tag{4.7}$$

with a variance of

$$V(\hat{\mu}_{na}^{st}) = \sum_{h=1}^{H} \sum_{l=1}^{L} (\frac{N_{hl}}{N})^2 \sigma_{hl}^2 (\frac{1}{n_{hl}} + \frac{1}{m_{hl}}). \tag{4.8}$$

Proof: Proof for unbiasedness of the mean estimator and for its variance is similar to that for theorem 4.3.

■

Stratified random sampling as above assumes that the strata sizes and the sampling frame for each stratum are available. However, situations do exist where the latter is difficult to obtain. In this situation it may be desirable to classify the units of a sample into strata after the sample is taken and to use a stratified estimate, even though the sample was selected by simple random sampling. This procedure

is termed *post-stratification.* This technique is useful where the published reports may provide clear indication of strata size, but the non-availability of strata frames makes sampling the units from different strata impractical. Mean estimation and its variance for sampling selection with and without replacement in post-stratified random sampling with a naive model are in the following two theorems.

Theorem 4.5 *Post-stratified Random Sampling without Replacement under the naive model*

In post-stratified random sampling without replacement under the naive model, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{pt} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nm_{hl}}\sum_{k\in R_{hl}} y_{hlk}, \tag{4.9}$$

with a conditional variance of

$$V(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}} - \frac{1}{N_{hl}})S_{hl}^2. \tag{4.10}$$

Proof: Suppose that sample size and response size are post-stratified as $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$ respectively. By lemmas 4.5 and 4.6,

$$\begin{aligned} E(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) &= E_I E_R(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) \\ &= E_I E_R \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm} \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\mu_{hl} \\ &= \mu. \end{aligned}$$

To prove equation 4.10, note that by using lemma 4.5, 4.6 and theorem 2.2

$$\begin{aligned} V_I E_R(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{n_{hl}} - \frac{1}{N_{hl}})S_{hl}^2, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})s_{hl}^2 \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})S_{hl}^2. \end{aligned}$$

Equation 4.10 follows by lemma 4.4.

■

Theorem 4.6 *Post-stratified Random Sampling with Replacement under the naive model*

In post-stratified random sampling with replacement under the naive model, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}_{na}^{pt} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nm_{hl}}\sum_{k\in R_{hl}} y_{hlk}, \tag{4.11}$$

with a conditional variance of

$$V(\hat{\mu}_{na}^{pt}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}}+\frac{1}{n_{hl}})\sigma_{hl}^2. \tag{4.12}$$

Proof: Proof for unbiasedness of the mean estimator and for its conditional variance is similar to that for theorem 4.5 but use the result of theorem 2.1.

■

Equal probability sampling is the procedure whereby each unit in the population has an equal chance of being included in the sample. However, when the units vary considerably in size of sampling units, equal probability sampling does not seem to be an appropriate procedure, since it does not take into account the possible importance of the size of the unit. Under such circumstances, selection of units with unequal probabilities may provide more efficient estimators than equal probability

sampling. In this scheme, the units are selected with probability proportional to a given measure of size. The size measure is the value of an auxiliary variable X, which is closely associated with the study variable Y. This type of sampling is known as *varying probability sampling* or *probability proportional to size (PPS) sampling.* For more details and notation used see section 2.3.3. The following four theorems deal with unequal probability sampling in simple and stratified random sampling both with and without replacement.

Theorem 4.7 *Random Equal Probability Sampling without Replacement under the naive model*

In random sampling with varying probabilities without replacement under the naive model, an unbiased estimator of the total τ is

$$\hat{\tau}^{srs}_{na,\pi ps} = \sum_{k\in R} \frac{ny_k}{m\pi_k}, \tag{4.13}$$

with a variance of

$$V(\hat{\tau}^{srs}_{na,\pi ps}) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N}\sum_{i\neq k}^{N} (\frac{\pi_{ki}-\pi_k\pi_i}{\pi_k\pi_i}) y_k y_i + E_I[n^2(1-f_1)\frac{s^2_{\tilde{y}_m}}{m}], \tag{4.14}$$

where $s^2_{\tilde{y}_m}$ is the response sample variance of $\frac{y_k}{\pi_k}$ and $f_1 = \frac{m}{n}$.

Proof: Using lemma 4.1 and theorem 2.8,

$$\begin{aligned} E(\hat{\tau}^{srs}_{na,\pi ps}) &= E_I E_R \frac{n}{m} \sum_{k=1}^{n} \frac{R_k y_k}{\pi_k} \\ &= E_I \sum_{k=1}^{n} \frac{y_k}{\pi_k} \\ &= \tau. \end{aligned}$$

To prove equation 4.14, note that by using lemmas 4.2, 4.3 and theorem 2.8

$$\begin{aligned} V_I E_R(\hat{\tau}^{srs}_{na,\pi ps}) &= V_I E_R[\frac{n}{m}\sum_{k=1}^{n} \frac{R_k y_k}{\pi_k}] \\ &= V_I[\sum_{k=1}^{n} \frac{y_k}{\pi_k}] \\ &= \sum_{k=1}^{N} (\frac{1-\pi_k}{\pi_k}) y_k^2 + \sum_{k=1}^{N}\sum_{i\neq k}^{N} (\frac{\pi_{ki}-\pi_k\pi_i}{\pi_k\pi_i}) y_k y_i, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{na,\pi ps}) &= E_I V_R[\frac{n}{m}\sum_{k=1}^{m}\frac{y_k}{\pi_k}] \\ &= E_I n^2 V_R(\tilde{\tau}_m) \\ &= E_I[n^2(1-f_1)\frac{s^2_{\tilde{y}_m}}{m}], \end{aligned}$$

where $\tilde{y}_k = \frac{y_k}{\pi_k}$, $f_1 = \frac{m}{n}$, $s^2_{\tilde{y}_m} = \frac{1}{m-1}\sum_{k=1}^{m}(\tilde{y}_k - \tilde{\tau}_m)^2$ and $\tilde{\tau}_m = \frac{1}{m}\sum_{k=1}^{m}\tilde{y}_k$ and equation 4.14 follows by the use of lemma 4.1.

■

Theorem 4.8 *Random Equal Probability Sampling with Replacement under the naive model*

In random sampling with varying probabilities with replacement under the naive model, an unbiased estimator of the total τ is

$$\hat{\tau}^{srs}_{na,pps} = \frac{1}{m}\sum_{k\in R}\frac{y_k}{p_k}, \tag{4.15}$$

with a variance of

$$V(\hat{\tau}^{srs}_{na,pps}) = \frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k} - \tau^2] + E_I(\frac{s^2_{\tilde{y}_m}}{m}), \tag{4.16}$$

where $s^2_{\tilde{y}}$ is the response sample variance of $\frac{y_k}{\pi_k}$.

Proof: An unbiased estimator of the population total is proved as in theorem 4.7 by using theorem 2.7 rather than 2.8.

To prove equation 4.16 note that by using lemmas 4.2, 4.3 and theorem 2.7

$$V_I E_R(\hat{\tau}^{srs}_{na,pps}) = \frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k} - \tau^2],$$

and

$$E_I V_R(\hat{\tau}^{srs}_{na,pps}) = E_I[\frac{s^2_{\tilde{y}_m}}{m}],$$

■

Theorem 4.9 *Stratified Equal Probability Sampling without Replacement under the naive model*

In stratified sampling with varying probabilities without replacement under the naive model, an unbiased estimator of the total τ is

$$\hat{\tau}^{st}_{na,\pi ps} = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{n_{hl}}{m_{hl}})\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}}, \tag{4.17}$$

with a variance of

$$V(\hat{\tau}^{st}_{na,\pi ps}) = \textstyle\sum_{h=1}^{H}\sum_{l=1}^{L}[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk}+$$

$$\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}] + E_I[\sum_{h=1}^{H}\sum_{l=1}^{L}n^2_{hl}(1-f_{hl1})\frac{s^2_{\bar{y}_{hlm}}}{m_{hl}}], \tag{4.18}$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$.

Proof: By the results from theorem 4.7, for each stratum, $E(\hat{\tau}^{srs}_{hl,na,\pi ps}) = \tau_{hl}$. Hence

$$\begin{aligned} E(\hat{\mu}^{st}_{na,\pi ps}) &= \sum_{h=1}^{H}\sum_{l=1}^{L}E(\hat{\tau}_{hlm}) \\ &= \sum_{h=1}^{H}\sum_{l=1}^{L}\tau_{hl} \\ &= \tau. \end{aligned}$$

Since the selections in different strata are independent and the variance in each stratum is by theorem 4.7, $\sum_{k=1}^{N_{hl}}\frac{1-\pi_{hlk}}{\pi_{hlk}}y^2_{hlk} + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}$ $+ E_I[n^2_{hl}(1-f_{hl1})\frac{s^2_{\bar{y}_{hlm}}}{m}]$. Equation 4.18 follows.

■

Theorem 4.10 *Stratified Equal Probability Sampling with Replacement under the naive model*

In stratified sampling with varying probabilities with replacement under the naive model, an unbiased estimator of the total τ is

$$\hat{\tau}^{st}_{na,pps} = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{1}{m_{hl}})\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.19}$$

with a variance of

$$V(\hat{\tau}^{st}_{na,pps}) = \sum_{h=1}^{H}\sum_{l=1}^{L}\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}] + E_I[\sum_{h=1}^{H}\sum_{l=1}^{L}\frac{s^2_{\bar{y}_{hl}}}{m_{hl}}]. \tag{4.20}$$

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that for theorem 4.9.

■

4.3.2 Weighting Adjustment Methods with the *RHG* Models

In simple random sampling with or without replacement scheme, a sample of size n is drawn from a population of size N in such a way that every possible sample of size n has the same chance of being selected. Suppose there are only m responses during data collection. Under *RHG* models, there are the four alternative methods of weighting class adjustment: sample-based, population-based, raking-ratio adjustment and general-based. For unequal probability sampling with replacement a bias-removal method is an alternative choice to use. This is a new adjustment method that I have developed. *Oh & Scheuren* (1983) give results of theorems for sample-based, population-based and raking ratio method in equal probability simple random sampling without replacement without proof. *Sarndal et al* (1992) give results of theorem for a sample-based adjustment method for unequal probability sampling without replacement without proof. I supply a proof for these cases and also for weighting adjustment in the *RHG* models with different sampling designs. The five *RHG* methods are presented separately in section 4.3.2.1-4.3.2.5. In section 4.3.2.1-4.3.2.4 theorem with equal probability simple random sampling without replacement is presented first followed by theorems with equal probability simple random sampling with replacement, unequal probability simple random sampling without and with replacement respectively. Theorems 4.14, 4.18, 4.22 and 4.26 for unequal probability sampling with replacement parallel Theorems 4.13, 4.17, 4.21

and 4.25 for sampling without replacement. However they are all under-estimate. These total estimators can be rescaled by using $\frac{1}{m_{hl}}$ to get a conditional unbiased estimator. These leads to the bias-removal method for unequal probability sampling with replacement that I have proposed. Section 4.3.2.5 presents the proof for this method .

4.3.2.1 Sample-based Adjustment Methods

Theorem 4.11 *Simple Random Sampling without Replacement under the sample-based adjustment method*

In simple random sampling without replacement, the sample-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_s^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hlm}, \tag{4.21}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2(\frac{1}{m_{hl}} - \frac{1}{N_{hl}})S_{hl}^2, \tag{4.22}$$

where $\mathbf{n} = (n_{11}, ..., n_{H_S L_S})$ and $\mathbf{m} = (m_{11}, ..., m_{H_S L_S})$.

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned} E(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hlm}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\mu_{hl} \\ &= \mu - \frac{1}{N}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\mu_{hl}-\mu)(N_{hl} - \frac{Nn_{hl}}{n}). \end{aligned}$$

To prove equation 4.22, note that by lemmas 4.5, 4.6 and theorem 2.2

$$\begin{aligned} V_I E_R(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2(\frac{1}{n_{hl}} - \frac{1}{N_{hl}})S_{hl}^2, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hlm}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})s_{hl}^2 \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})S_{hl}^2. \end{aligned}$$

Equation 4.22 follows by lemma 4.4.

■

Theorem 4.12 *Simple Random Sampling with Replacement under the sample-based adjustment method*

In simple random sampling with replacement, the sample-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_s^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\mu}_{hlm}, \tag{4.23}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\sigma_{hl}^2(\frac{1}{n_{hl}}+\frac{1}{m_{hl}}). \tag{4.24}$$

Proof: The conditional biased estimator $\hat{\mu}_s^{srs}$ is proved as in theorem 4.11.

To prove equation 4.24 note that by lemmas 4.5, 4.6 and theorem 2.1

$$V_I E_R(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\frac{\sigma_{hl}^2}{n_{hl}},$$

and

$$E_I V_R(\hat{\mu}_s^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\frac{\sigma_{hl}^2}{m_{hl}}.$$

Equation 4.24 follows by lemma 4.4. ■

Theorem 4.13 *Random Unequal Probability Sampling without Replacement under the sample-based adjustment method*

In random sampling with varying probabilities without replacement, the sample-based weighting adjustment estimator of the total τ,

$$\hat{\tau}^{srs}_{s,\pi ps} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{m_{hl}}\sum_{k\in R_{hl}}\frac{y_{hlk}}{\pi_{hlk}}, \tag{4.25}$$

is a conditional unbiased estimator with a conditional variance of

$$V(\hat{\tau}^{srs}_{s,\pi ps}|\mathbf{n},\mathbf{m}) = \textstyle\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk}+$$

$$\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}] + E_I[\sum_{h=1}^{H}\sum_{l=1}^{L}(1-f_{hl1})n^2_{hl}\frac{s^2_{\tilde{y}_{hl}}}{m_{hl}}], \tag{4.26}$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$, $s^2_{\tilde{y}_{hlm}} = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk}-\tilde{\tau}_{hlm})^2$, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$, $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}^{srs}_{s,\pi ps}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{m_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}]\\
&= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\hat{\tau}_{hl}\\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl}\\
&= \tau.
\end{aligned}$$

To prove equation 4.26, note that by lemmas 4.5, 4.6 and theorem 2.8

$$\begin{aligned}
V_I E_R(\hat{\tau}^{srs}_{s,\pi ps}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}]\\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk} + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}],
\end{aligned}$$

and

$$\begin{aligned}
E_I V_R(\hat{\tau}^{srs}_{s,\pi ps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{m_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}}]\\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}n^2_{hl}V_R(\tilde{\tau}_{hlm})]\\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}n^2_{hl}(1-f_{hl1})\frac{s^2_{\tilde{y}_{hl}}}{m_{hl}}].
\end{aligned}$$

Equation 4.26 follows by lemma 4.4. ∎

Theorem 4.14 *Random Unequal Probability Sampling with Replacement under the sample-based adjustment method*

In random sampling with varying probabilities with replacement, the sample-based weighting adjustment estimator of the total τ,

$$\hat{\tau}_{s,pps}^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{nm_{hl}}\sum_{k\in R_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.27}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}_{s,pps}^{srs}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}^2}{p_{hlk}} - \tau_{hl}^2] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\frac{s_{\tilde{y}_{hl}}^2}{m_{hl}}]. \tag{4.28}$$

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}_{s,pps}^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{nm_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\
&= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\hat{\tau}_{hl} \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{n}\tau_{hl} \\
&< \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl} \\
&< \tau.
\end{aligned}$$

To prove equation 4.28, note that by lemmas 4.5, 4.6 and theorem 2.7

$$\begin{aligned}
V_I E_R(\hat{\tau}_{s,pps}^{srs}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{n}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2 V_I[\frac{1}{n_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}^2}{p_{hlk}} - \tau_{hl}^2],
\end{aligned}$$

and

$$E_I V_R(\hat{\tau}_{S,pps}^{srs}|\mathbf{n},\mathbf{m}) = E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{n_{hl}}{nm_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}]$$

$$
\begin{aligned}
&= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2 V_R(\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}) \\
&= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2 V(\tilde{\tau}_{hlm}) \\
&= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{n_{hl}}{n})^2 \frac{s^2_{\tilde{y}_{hl}}}{m_{hl}}.
\end{aligned}
$$

Equation 4.28 follows by lemma 4.4.

■

4.3.2.2 Population-based Adjustment Methods

Theorem 4.15 *Simple Random Sampling without Replacement under the population-based adjustment method*

In simple random sampling without replacement, the population-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_p^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm}, \tag{4.29}$$

is a conditional unbiased estimator with a conditional variance of

$$V(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}} - \frac{1}{N_{hl}})S^2_{hl}, \tag{4.30}$$

where $\mathbf{n} = (n_{11}, ..., n_{H_S L_S})$ and $\mathbf{m} = (m_{11}, ..., m_{H_S L_S})$.

Proof: By lemmas 4.5 and 4.6,

$$
\begin{aligned}
E(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm} \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\mu_{hl} \\
&= \mu.
\end{aligned}
$$

To prove equation 4.30, note that by using lemmas 4.5, 4.6 and theorem 2.2

$$
\begin{aligned}
V_I E_R(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hl}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{n_{hl}} - \frac{1}{N_{hl}})S^2_{hl},
\end{aligned}
$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})s_{hl}^2 \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})S_{hl}^2. \end{aligned}$$

Equation 4.30 follows by lemma 4.4.

■

Theorem 4.16 *Simple Random Sampling with Replacement under the population-based adjustment method*

In simple random sampling with replacement, the population-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_p^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\mu}_{hlm}, \tag{4.31}$$

is a conditional unbiased estimator with a conditional variance of

$$V(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\sigma_{hl}^2(\frac{1}{n_{hl}}+\frac{1}{m_{hl}}). \tag{4.32}$$

Proof: The conditional unbiased estimator $\hat{\mu}_p^{srs}$ is proved as in theorem 4.15.

To prove equation 4.32, note that by using lemmas 4.5, 4.6 and theorem 2.1

$$V_I E_R(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{\sigma_{hl}^2}{n_{hl}},$$

and

$$E_I V_R(\hat{\mu}_p^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{\sigma_{hl}^2}{m_{hl}}.$$

Equation 4.32 follows by lemma 4.4.

■

Theorem 4.17 *Random Unequal Probability Sampling without Replacement under the population-based adjustment method*

In random sampling with varying probabilities without replacement, the population-based weighting adjustment estimator of the total τ,

$$\hat{\tau}_{p,\pi ps}^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}n}{m_{hl}N}\sum_{k\in R_{hl}}\frac{y_{hlk}}{\pi_{hlk}}, \tag{4.33}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}_{p,\pi ps}^{srs}|\mathbf{n},\mathbf{m}) = \textstyle\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{nN_{hl}}{Nn_{hl}})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y_{hlk}^2+$$

$$\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(1-f_{hl1})(\frac{N_{hl}n}{N})^2\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}}], \tag{4.34}$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$, $s_{\tilde{y}_{hlm}}^2 = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk}-\tilde{\tau}_{hlm})^2$, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$, $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}_{p,\pi ps}^{srs}|\mathbf{n},\mathbf{m}) &= E_IE_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}n}{m_{hl}N}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}]\\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{n_{hl}N})\hat{\tau}_{hl}]\\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{n_{hl}N})\tau_{hl}\\
&\neq \tau.
\end{aligned}$$

To prove equation 4.34, note that by lemmas 4.5, 4.6 and theorem 2.8

$$\begin{aligned}
V_IE_R(\hat{\tau}_{p,\pi ps}^{srs}|\mathbf{n},\mathbf{m}) &= V_IE_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}n}{m_{hl}N}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}]\\
&= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}n}{n_{hl}N}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}]\\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{n_{hl}N})^2V_I[\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}]\\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{n_{hl}N})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y_{hlk}^2+\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}],
\end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{p,\pi ps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{m_{hl}N})\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{N})^2 V_R(\tilde{\tau}_{hlm})] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}n}{N})^2(1-f_{hl1})\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}]. \end{aligned}$$

Equation 4.34 follows by lemma 4.4.

■

Theorem 4.18 *Random Unequal Probability Sampling with Replacement under the population-based adjustment method*

In random sampling with varying probabilities with replacement, the population-based weighting adjustment estimator of the total τ,

$$\hat{\tau}^{srs}_{p,pps} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nm_{hl}}\sum_{k\in R_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.35}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}^{srs}_{p,pps}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}]. \tag{4.36}$$

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned} E(\hat{\tau}^{srs}_{p,pps}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nm_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\hat{\tau}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N}\tau_{hl}) \\ &< \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl} \\ &< \tau. \end{aligned}$$

To prove equation 4.36, note that by lemmas 4.5, 4.6 and theorem 2.7

$$\begin{aligned} V_I E_R(\hat{\tau}^{srs}_{p,pps}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nn_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V_I[\frac{1}{n_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}], \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{p,pps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nm_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V_R(\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}) \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V(\tilde{\tau}_{hlm}) \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}. \end{aligned}$$

Equation 4.36 follows by lemma 4.4.

■

4.3.2.3 Raking Ratio Adjustment Methods

Theorem 4.19 *Simple Random Sampling without Replacement under the raking ratio adjustment method*

In simple random sampling without replacement, the raking ratio weighting adjustment estimator of the mean μ,

$$\hat{\mu}^{srs}_r = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{N}\hat{\mu}_{hlm}, \tag{4.37}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}^{srs}_r|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2(\frac{1}{m_{hl}} - \frac{1}{N_{hl}})S^2_{hl}, \tag{4.38}$$

where $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: By using lemmas 4.5 and 4.6

$$\begin{aligned} E(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R \sum_{h=1}^{H_S}\sum_{l=1}^{L_S} \frac{N_{hl}^*}{N}\hat{\mu}_{hlm} \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S} \frac{N_{hl}^*}{N}\mu_{hl} \\ &= \mu - \frac{1}{N}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\mu_{hl}-\mu)(N_{hl}-N_{hl}^*). \end{aligned}$$

To prove equation 4.38, note that by lemmas 4.5, 4.6 and theorem 2.2

$$\begin{aligned} V_I E_R(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} \frac{N_{hl}^*}{N}\hat{\mu}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2(\frac{1}{n_{hl}}-\frac{1}{N_{hl}})S_{hl}^2, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} \frac{N_{hl}^*}{N}\hat{\mu}_{hlm}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})s_{hl}^2 \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})S_{hl}^2. \end{aligned}$$

Equation 4.38 follows by lemma 4.4.

■

Theorem 4.20 *Simple Random Sampling with Replacement under the raking ratio adjustment method*

In simple random sampling with replacement, the raking ratio weighting adjustment estimator of the mean μ,

$$\hat{\mu}_r^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S} \frac{N_{hl}^*}{N}\hat{\mu}_{hlm}, \tag{4.39}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2\sigma_{hl}^2(\frac{1}{n_{hl}}+\frac{1}{m_{hl}}). \tag{4.40}$$

Proof: The conditional biased estimator $\hat{\mu}_r^{srs}$ is proved as in theorem 4.19.

To prove equation 4.40, note that by lemmas 4.5, 4.6 and theorem 2.1

$$V_I E_R(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2\frac{\sigma_{hl}^2}{n_{hl}},$$

and

$$E_I V_R(\hat{\mu}_r^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^*}{N})^2\frac{\sigma_{hl}^2}{m_{hl}}.$$

Equation 4.40 then follows by lemma 4.4.

■

Theorem 4.21 *Random Unequal Probability Sampling without Replacement under the raking ratio adjustment method*

In random sampling with varying probabilities without replacement, the raking ratio weighting adjustment estimator of the total τ,

$$\hat{\tau}_{r,\pi ps}^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}^* n}{m_{hl}N}\sum_{k\in R_{hl}}\frac{y_{hlk}}{\pi_{hlk}}, \tag{4.41}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}_{r,\pi ps}^{srs}|\mathbf{n},\mathbf{m}) = \textstyle\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{nN_{hl}^*}{Nn_{hl}})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y_{hlk}^2 +$$

$$\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(1-f_{hl1})(\frac{N_{hl}^* n}{N})^2\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}}], \tag{4.42}$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$, $s_{\tilde{y}_{hlm}}^2 = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk}-\tilde{\tau}_{hlm})^2$, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$, $\mathbf{n} = (n_{11},...,n_{H_S L_S})$ and $\mathbf{m} = (m_{11},...,m_{H_S L_S})$.

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned} E(\hat{\tau}_{r,\pi ps}^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}^* n}{m_{hl}N}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^* n}{n_{hl}N})\hat{\tau}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}^* n}{n_{hl}N})\tau_{hl} \\ &\neq \tau. \end{aligned}$$

To prove equation 4.42, note that by lemmas 4.5, 4.6 and theorem 2.8

$$\begin{aligned} V_I E_R(\hat{\tau}^{srs}_{r,\pi ps}|\mathbf{n},\mathbf{m}) &= V_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}n}{m_{hl}N}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}] \\ &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}n}{n_{hl}N}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}n}{n_{hl}N})^2 V_I[\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}n}{n_{hl}N})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk} + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}], \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{r,\pi ps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}n}{m_{hl}N})\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}n}{N})^2 V_R(\tilde{\tau}_{hlm})] \\ &= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}n}{N})^2(1-f_{hl1})\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}]. \end{aligned}$$

Equation 4.42 follows by lemma 4.4.

■

Theorem 4.22 *Random Unequal Probability Sampling with Replacement under the raking ratio adjustment method*

In random sampling with varying probabilities with replacement, the raking ratio weighting adjustment estimator of the total τ,

$$\hat{\tau}^{srs}_{r,pps} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{Nm_{hl}}\sum_{k\in R_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.43}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}^{srs}_{r,pps}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}]. \tag{4.44}$$

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}^{srs}_{R,pps}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{Nm_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{N}\hat{\tau}_{hl}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N}\tau_{hl}) \\
&< \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl} \\
&< \tau.
\end{aligned}$$

To prove equation 4.44, note that by lemmas 4.5, 4.6 and theorem 2.7

$$\begin{aligned}
V_I E_R(\hat{\tau}^{srs}_{r,pps}|\mathbf{n},\mathbf{m}) &= V_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{Nm_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\
&= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{Nn_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2 V_I[\frac{1}{n_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}],
\end{aligned}$$

and

$$\begin{aligned}
E_I V_R(\hat{\tau}^{srs}_{r,pps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N^*_{hl}}{Nm_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2 V_R(\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}) \\
&= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2 V(\tilde{\tau}_{hlm}) \\
&= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N^*_{hl}}{N})^2\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}.
\end{aligned}$$

Equation 4.44 follows by lemma 4.4.

■

4.3.2.4 General-based Adjustment Methods

Theorem 4.23 *Simple Random Sampling without Replacement under the general-based adjustment method*

In simple random sampling without replacement, the general-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_g^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hlm}, \tag{4.45}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2(\frac{1}{m_{hl}}-\frac{1}{N_{hl}})S_{hl}^2, \tag{4.46}$$

where $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: By lemmas 4.5 and 4.6

$$\begin{aligned} E(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hlm} \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hl} \\ &= \mu - \frac{1}{N}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\mu_{hl}-\mu)(N_{hl}-\frac{N}{m}m_{hl}). \end{aligned}$$

To prove equation 4.46, note that by lemmas 4.5, 4.6 and theorem 2.2

$$\begin{aligned} V_I E_R(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hl}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2(\frac{1}{n_{hl}}-\frac{1}{N_{hl}})S_{hl}^2, \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hlm}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})s_{hl}^2 \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2(\frac{1}{m_{hl}}-\frac{1}{n_{hl}})S_{hl}^2. \end{aligned}$$

Equation 4.46 follows by lemma 4.4.

■

Theorem 4.24 *Simple Random Sampling with Replacement under the general-based adjustment method*

In simple random sampling with replacement, the general-based weighting adjustment estimator of the mean μ,

$$\hat{\mu}_g^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\mu}_{hlm}, \tag{4.47}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2\sigma_{hl}^2(\frac{1}{n_{hl}}+\frac{1}{m_{hl}}), \tag{4.48}$$

where $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: The conditional biased estimator $\hat{\mu}_g^{srs}$ is proved as in theorem 4.23.

To prove equation 4.48, note that by lemmas 4.5, 4.6 and theorem 2.2

$$V_IE_R(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2(\frac{\sigma_{hl}^2}{n_{hl}}),$$

and

$$E_IV_R(\hat{\mu}_g^{srs}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2\frac{\sigma_{hl}^2}{m_{hl}}.$$

Equation 4.48 follows by lemma 4.4.

■

Theorem 4.25 *Random Unequal Probability Sampling without Replacement under the general-based adjustment method*

In random sampling with varying probabilities without replacement, the general-based weighting adjustment estimator of the total τ,

$$\hat{\tau}_{g,\pi ps}^{srs} = \frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k\in R_{hl}}\frac{y_{hlk}}{\pi_{hlk}}, \tag{4.49}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}^{srs}_{g,\pi ps}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{nm_{hl}}{mn_{hl}})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk}+$$

$$\sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(1-f_{hl1})(\frac{nm_{hl}}{m})^2\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}], \qquad (4.50)$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$, $s^2_{\tilde{y}_{hlm}} = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk}-\tilde{\tau}_{hlm})^2$, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$, $\mathbf{n} = (n_{11},...,n_{H_SL_S})$ and $\mathbf{m} = (m_{11},...,m_{H_SL_S})$.

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}^{srs}_{g,\pi ps}|\mathbf{n},\mathbf{m}) &= E_IE_R[\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}] \\
&= E_I\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{n_{hl}}\hat{\tau}_{hl} \\
&= \frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{n_{hl}}\tau_{hl} \\
&\neq \tau.
\end{aligned}$$

To prove equation 4.50, note that by lemmas 4.5, 4.6 and theorem 2.8

$$\begin{aligned}
V_IE_R(\hat{\tau}^{srs}_{g,\pi ps}|\mathbf{n},\mathbf{m}) &= V_I[\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{\pi_{hlk}}] \\
&= V_I[\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{n_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}] \\
&= (\frac{n}{m})^2\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{n_{hl}})^2V_I[\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{\pi_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{nm_{hl}}{mn_{hl}})^2[\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y^2_{hlk} + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N_{hl}}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli}],
\end{aligned}$$

and

$$\begin{aligned}
E_IV_R(\hat{\tau}^{srs}_{g,\pi ps}|\mathbf{n},\mathbf{m}) &= E_IV_R[\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}}] \\
&= E_IV_R[\frac{n}{m}\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}m_{hl}\tilde{\tau}_{hlm}] \\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{nm_{hl}}{m})^2(1-f_{hl1})\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}],
\end{aligned}$$

where $f_{hl1} = \frac{m_{hl}}{n_{hl}}$, $\tilde{y}_{hlk} = \frac{y_{hlk}}{\pi_{hlk}}$, $s^2_{\tilde{y}_{hlm}} = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk}-\tilde{\tau}_{hlm})^2$ and $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$.

Equation 4.50 follows by lemma 4.4.

■

Theorem 4.26 *Random Unequal Probability Sampling with Replacement under the general-based adjustment method*

In random sampling with varying probabilities with replacement, the general-based weighting adjustment estimator of the total τ,

$$\hat{\tau}_{g,pps}^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m}\sum_{k\in R_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.51}$$

is a conditional biased estimator with a conditional variance of

$$V(\hat{\tau}_{g,pps}^{srs}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}^2}{p_{hlk}} - \tau_{hl}^2] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}}]. \tag{4.52}$$

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned}
E(\hat{\tau}_{g,pps}^{srs}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\
&= E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{m_{hl}}{m}\hat{\tau}_{hl}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m}\tau_{hl}) \\
&< \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl} \\
&< \tau.
\end{aligned}$$

To prove equation 4.52, note that by lemmas 4.5, 4.6 and theorem 2.7

$$\begin{aligned}
V_I E_R(\hat{\tau}_{g,pps}^{srs}|\mathbf{n},\mathbf{m}) &= V_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\
&= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{mn_{hl}})\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2 V_I[\frac{1}{n_{hl}}\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\
&= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}^2}{p_{hlk}} - \tau_{hl}^2],
\end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{g,pps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2 V_R(\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}) \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2 V_R(\tilde{\tau}_{hlm}) \\ &= E_I \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{m_{hl}}{m})^2 \frac{s^2_{\tilde{y}_{hl}}}{m_{hl}}, \end{aligned}$$

where $\tilde{y}_{hlk} = \frac{y_{hlk}}{p_{hlk}}$, $s^2_{\tilde{y}_{hl}} = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk} - \tilde{\tau}_{hlm})^2$ and $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}$.

Equation 4.52 follows by lemma 4.4.

■

Theorems 4.14, 4.18, 4.22 and 4.26 relate to *RHG* models under unequal probability random sampling with replacement. As shown these total estimators are all under-estimates. Theorem 4.27 below defines an estimator I call the *bias-removal* method which gives a conditional unbiased estimator in all these cases.

4.3.2.5 Bias-removal Methods

Theorem 4.27 *Random Equal Probability Sampling with Replacement under the bias-removal method*

In random sampling with varying probabilities with replacement under the bias-removal method, a conditional unbiased estimator of the total τ is

$$\hat{\tau}^{srs}_{u,pps} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m_{hl}}\sum_{k\in R_{hl}}\frac{y_{hlk}}{p_{hlk}}, \tag{4.53}$$

with a conditional variance of

$$V(\hat{\tau}^{srs}_{u,pps}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}] + E_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}}]. \tag{4.54}$$

Proof: By lemmas 4.5 and 4.6,

$$\begin{aligned} E(\hat{\tau}^{srs}_{u,pps}|\mathbf{n},\mathbf{m}) &= E_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m_{hl}}\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\hat{\tau}_{hl} \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\tau_{hl} \\ &= \tau \end{aligned}$$

To prove equation 4.54, note that by lemmas 4.5, 4.6 and theorem 2.7

$$\begin{aligned} V_I E_R(\hat{\tau}^{srs}_{u,pps}|\mathbf{n},\mathbf{m}) &= V_I E_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{1}{m_{hl}})\sum_{k=1}^{n_{hl}}\frac{R_{hlk}y_{hlk}}{p_{hlk}}] \\ &= V_I[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{1}{n_{hl}})\sum_{k=1}^{n_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}], \end{aligned}$$

and

$$\begin{aligned} E_I V_R(\hat{\tau}^{srs}_{u,pps}|\mathbf{n},\mathbf{m}) &= E_I V_R[\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}}] \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}V_R(\frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\tilde{y}_{hlk}) \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}V(\tilde{\tau}_{hlm}) \\ &= E_I\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{s^2_{\tilde{y}_{hl}}}{m_{hl}}. \end{aligned}$$

Equation 4.54 follows by lemma 4.4.

■

Chapter 5

Imputation Methods

This chapter focuses on imputation methods to deal with nonresponse problems after data have been collected. Weighting adjustment methods, the alternative way to deal these nonresponse problems after data have been collected, was presented in chapter 4. Imputation methods are introduced in section 5.1. Section 5.2 introduces single imputation methods namely random, sequential and stochastic regression methods. Section 5.3 presents multiple imputation methods.

5.1 Overview

Weighting adjustment for nonresponse discussed in chapter 4 makes a strong assumption; in each weighting cell, the respondents and nonrespondents are assumed to come from the same population. This assumption never exactly describes the true state of affairs because, generally, respondents and nonrespondents tend to behave differently and this leads to bias. Weights may improve many of the estimates, but they rarely eliminate all nonresponse bias. The extent of the bias depends on the type of response model used in the analysis. If weighting adjustments are made, statisticians should always state the assumed response model and give evidence to

justify it. For more details on different response model mechanisms see *Lohr* (1999).

Imputation is commonly used to assign values to the missing items due to non-response. The replacement value is often from another sample unit in the survey which is similar to the item nonresponse. Imputation procedures are used not only to reduce the nonresponse bias but to produce a *complete rectangular data matrix*, often called a *clean data matrix*, for standard complete-data methods of analysis (*Lohr*, 1999).

Generally nonresponse can be one of two types: (i) item nonresponse where some items in a sample unit are missing or (ii) unit nonresponse where all items in a sample unit are missing. This leads to two principal types of imputation (*Sarndal et al*, 1992):

i) *Imputation for item nonresponse only*: Imputed values provide missing values corresponding to element k in the item nonresponse set , $r_u - r_c$. The r_u is called the unit response set where there is response one or more items. The r_c is called the complete response set which is composed of the elements having responded to all items. Weighting adjustment is then applied to compensate for the unit nonresponse and the unit nonresponse set $s - r_u$ is discarded.

ii) *Imputation for item nonresponse as well as for unit nonresponse*: Imputed values provide missing values corresponding to missing elements in the partial response set $s - r_c$, where s is the sample set. No weighting adjustment is applied. Estimates are computed using genuine, as well as imputed, values.

Thus a complete rectangular matrix is the result of imputation, in both cases (i) and (ii). The matrix is of dimension $n_{r_u} \times q$ in case (i) and $n_s \times q$ in case (ii) where q is the number of items in the sampling unit.

Many researcher ignore nonresponse. The effect of this is to obtain a matrix that uses the observed y data from elements in r_c only. By treating r_c as a reduced

unit response set, a researcher would then apply the usual techniques for unit non-response. However, this may bias the estimator and the estimated variance may be larger than the desired variance from the survey planning.

A number of imputation techniques have been developed, as discussed by *Sande* (1982,1983), *Bailar et al* (1978), *Ford* (1983), *Kalton & Kasprzyk* (1986), Little & Rubin (1987), *Little* (1988) , *Lohr* (1999), *Lessler & Kalsbeek* (1992), *Levy & lemeshow* (1999), *Armitage & Colton* (1999), *Govindarajulu* (1999), *Jinn et al* (1989a, 1989b). Eleven imputation procedures selected from these are briefly discussed below. These are deductive, overall mean, cell mean imputation, stochastic regression, substitution, cold-deck and several hot-deck imputation methods.

i) *Deductive Imputation*: This method refers to those instances, rare in practice, where a missing value can be filled with a good prediction $\tilde{y}_{ki} = y_{ki}$, attained by logical conclusion. Such deduction is sometimes used in longitudinal surveys.

ii) *Overall Mean Imputation*: This method assigns the overall respondent mean to all missing responses. Unless the nonresponse is negligible, or unless a modified variance estimator is used, the method may easily lead to seriously understated variance estimates and to invalid confidence intervals.

iii) *Cell Mean Imputation*: Respondents are divided into classes based on known variables, as in weighting class adjustments. The average of the values for the responding units in cell hl, $\hat{\mu}_{hlm}$ is substituted for each missing value in the cell. Cell mean imputation assumes that missing items are missing completely at random within cells. However this method also fail to reflect the variability of the nonrespondents.

Improvement on the mean imputation methods is sought by creating a more authentic variability in the imputed values. Substitution, stochastic regression, *cold-deck* or *hot-deck* methods are commonly used.

iv) *Substitution*: Sometimes interviewers are allowed to choose a substitute while in the field. In case a substitution is used, it should be reported. For example, if the household selected for the sample is not at home, the next household is used. Substitution may help reduce some nonresponse bias because the household next door may be more similar to the nonresponding household than would be a household selected at random from the population. Effect on selection probabilities is however important in this case. Houses next to potential nonrespondents have a higher selection probability.

v) *Regression Imputation*: This method predicts the missing value by using a regression of the item of the interest based on all the observed cases. A variation is stochastic regression imputation, in which the missing value is replaced by the predicted value from the regression model with an added residual terms which is used to avoid underestimatation of variance. There are several ways in which the residual may be obtained, e.g., randomly generated error term from a normal population with zero mean and variance of the regression residual. For more details on adding residuals see *Kalton* (1983) or *Govindarajuru* (1999).

vi) *Cold-deck Imputation*: This procedure uses imputations based on other sources than the current survey, for example, earlier surveys or historical data.

Hot-deck Imputation: Missing responses are replaced by values selected from respondents in the current survey in various ways. Several of these hot-deck imputation methods are given below:

vii) *Random Overall Imputation*: A respondent is chosen at random from the total respondent sample, and the selected respondent's value is assigned to the nonrespondent. This method is the simplest form of hot-deck imputation. Usually,

to preserve any multivariate relationships, values from the same donor are used for all missing items of the nonrespondent.

viii) *Random Imputation Within Class*: In this hot-deck method, a respondent is chosen at random within an imputation class, and the nonrespondent missing values are replaced by those of the selected respondent.

ix) *Sequential Imputation*: When a missing value is spotted on a certain item, a donor is identified by backtracking through the data file to the nearest element that shows a response value for the item that is in the same imputation class as the recipient. The procedure starts with a cold-deck value if the first unit is missing in each imputation class. This method has the advantage that a single pass through the data file is sufficient to complete the imputation procedure. One draw-back of this nonrandom procedure is that it often leads to multiple uses of donors. The ways of avoiding this problem are the use of multiple cold-deck values in registers that are rotated or the use of hierarchical sequential imputation.

x) *Hierarchical sequential Imputation*: This procedure sorts respondents and non-respondents into a large number of imputation classes from a detailed categorisation of a set of auxiliary variables. Nonrespondents are then matched with respondents on a hierarchical basis, in the sense that if a match cannot be made in the initial imputation class, classes are collapsed and the match is made at the lower level of detail.

xi) *Distance Function Matching* or *Nearest-Neighbour Imputation*: In this method a distance measure between observations is defined in terms of known auxiliary variable values. The value of a respondent which is closest to the sample unit with the missing item is used to impute a missing data. Various forms of distance functions have been proposed (e.g. *Sande*, 1979; *Vacek & Ashikago*,

1980) and the function can be constructed to reduce the multiple use of donors by incorporating a penalty for each use (*Colledge et al*, 1978).

The methods discussed above called *single imputation* methods involve replacing each missing value by a single imputed value. There are two major attractive features of this practice. Firstly, standard complete-data methods of analysis can be used in the resulting "clean data matrix". Secondly, in the context of public-use data bases, substantial effort is often required to create sensible imputations. With single imputation this need be carried out only once, by the data producer.

However, there is a serious disadvantage of single imputation methods; the single value being imputed can reflect neither sampling variability about the actual value when one model for nonresponse is used nor additional uncertainty when more than one model is considered (*Rubin*, 1987). To correct for this disadvantage and retain the virtues of single imputation, *multiple imputation* has been developed by *Rubin* (1987). The idea behind multiple imputation is that for each missing value, $M \geq 2$ different estimates are imputed. Typically, the same stochastic model is used for each imputation creating different clean data sets. Each of the M data sets is analysed as if no imputation had been done. The different results give the analyst a measure of the additional variance due to the imputation. When different models of nonresponse are used, multiple imputation can give an idea of the sensitivity of the results to particular nonresponse models. More details on implementing multiple imputation can be found in *Rubin* (1987,1996), *Bernard et al* (1998), *Rao* (1996) and *Fay* (1996).

Imputation methods are discussed in the two sections. Section 5.2 describes single imputation theorems. Section 5.3 presents multiple imputation procedures without stating the theorem.

5.2 Single Imputation

In this section, random imputation theorems are presented in section 5.2.1. Quasi-randomisation (see section 2.6.4.1) is assumed. Theorems about sequential imputation and stochastic regression imputation are presented in section 5.2.2 and 5.2.3 respectively. In these three sections the theorems on post-stratified random sampling under the naive model are not stated as these theorems are essential the same as those for simple random sampling under the *RHG* model. If nonresponse is greater than response then the theorems on random imputation on simple random sampling without replacement are not applicable. Sequential imputation theorems are only used for equal probability sampling with replacement.

Before section 5.2.1-5.2.3 are presented the following three lemmas are proved as these are used in the proof of the theorems in the random, sequential and stochastic regression cases.

In these lemmas and following theorems, notations for marginal distribution, joint distribution, expectation and variance in chapter 4 are used but where the additional $\mathbf{H}$ is sometimes used as subscript distribution and expectation where $\mathbf{H}$ is the number of times y is used as a substitute for a missing value of $\mathbf{Y}$, e.g., $f_H(.)$ is the marginal distribution of $\mathbf{H}$ given $\mathbf{I}, \mathbf{RY}, \mathbf{X}, \mathbf{n}, \mathbf{m}$.

Lemma 5.1 *Expectation and Variance in Random Imputation with Quasi-randomisation*

The expectation of an unbiased estimator $\hat{\theta}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ and $\mathbf{m}$ in random imputation with quasi-randomisation theory is

$$E(\hat{\theta}) = E(\hat{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}) = E_I E_R E_H(\hat{\theta}).$$

The variance of estimator $\hat{\theta}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}$ and $\mathbf{m}$ in random imputation with quasi-randomisation theory is

$$V(\hat{\theta}) = V(\hat{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}) = V_I E_R E_H(\hat{\theta}) + E_I V_R E_H(\hat{\theta}) + E_I E_R V_H(\hat{\theta}).$$

Proof: Proof is similar to that of lemma 4.1 with an additional random imputation step.

■

Lemma 5.2 *Expectation for Variance with Response in the Sample*

The expected variance with response in the sample is the sample variance for sampling with or without replacement:

$E(s_m^2) = s^2$, where $s_m^2 = \frac{1}{m-1}\sum_{k\in R}(y_k - \hat{\mu}_m)^2$, $s^2 = \frac{1}{n-1}\sum_{k=1}^{n}(y_k - \hat{\mu})^2$, $\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} y_k$ where R is the response indicator in the sample and m is a response size. *Proof:* Assume that $MCAR$ is used. Let R_k be the k^{th} response indicator in the sample which has Bernoulli distribution with probability of response $p(R_k = 1|\mathbf{I}, n, m) = \frac{m}{n}$.

Thus, $E(R_k) = \frac{m}{n}$, $E(R_k^2) = \frac{m}{n}$ and $E(R_k R_i) = \frac{m(m-1)}{n(n-1)}$.

$s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k - \hat{\mu}_m)^2$ can be rewritten as

$$\begin{aligned} s_m^2 &= \frac{1}{m-1}[\sum_{k=1}^{m} y_k^2 - \frac{1}{m}(\sum_{k=1}^{m} y_k)^2] \\ &= \frac{1}{m-1}[\sum_{k=1}^{n} R_k y_k^2 - \frac{1}{m}(\sum_{k=1}^{n} R_k y_k)^2]. \end{aligned}$$

Then the expectation of s_m^2 is

$$\begin{aligned} E(s_m^2) &= E\frac{1}{m-1}[\sum_{k=1}^{n} R_k y_k^2 - \frac{1}{m}(\sum_{k=1}^{n} R_k y_k)^2] \\ &= \frac{1}{m-1}E[\sum_{k=1}^{n} R_k y_k^2 - \frac{1}{m}(\sum_{k=1}^{n} R_k^2 y_k^2 + \sum_{k=1}^{n}\sum_{i\neq k}^{n} R_k R_i y_k y_i)] \\ &= \frac{1}{m-1}[\sum_{k=1}^{n}\frac{m}{n} y_k^2 - \frac{1}{m}\{\frac{m}{n}\sum_{k=1}^{n} y_k^2 + \frac{m(m-1)}{n(n-1)}\sum_{k=1}^{n}\sum_{i\neq k}^{n} y_k y_i\}] \\ &= \frac{1}{n(m-1)}[m\sum_{k=1}^{n} y_k^2 - \sum_{k=1}^{n} y_k^2 - \frac{(m-1)}{(n-1)}\sum_{k=1}^{n}\sum_{i\neq k}^{n} y_k y_i] \\ &= \frac{1}{n}[\sum_{k=1}^{n} y_k^2 - \frac{1}{n-1}\sum_{k=1}^{n}\sum_{i\neq k}^{n} y_k y_i] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n(n-1)}[n\sum_{k=1}^{n} y_k^2 - (\sum_{k=1}^{n} y_k^2 + \sum_{k=1}^{n}\sum_{i\neq k}^{n} y_k y_i)] \\
&= \frac{1}{n-1}[\sum_{k=1}^{n} y_k^2 - \frac{1}{n}(\sum_{k=1}^{n} y_k)^2] \\
&= s^2.
\end{aligned}
$$

■

Lemma 5.3 *Expectation and Variance for Post-stratified Sample in Random Imputation with Quasi-randomisation*

The expectation of an unbiased estimator $\hat{\theta}$ given $\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}$ for post-stratified sample in quasi-randomisation theory is

$$E(\hat{\theta}|\mathbf{n}, \mathbf{m}) = E(\hat{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{n}, \mathbf{m}) = E_I E_R E_H(\hat{\theta}|\mathbf{n}, \mathbf{m}).$$

The variance of an estimator $\hat{\theta}$ given $Y, X, \mathbf{n}, \mathbf{m}$ for post-stratified sample in quasi-randomisation theory is

$$V(\hat{\theta}|\mathbf{n}, \mathbf{m}) = V(\hat{\theta}|Y, X, \mathbf{n}, \mathbf{m}) =$$
$$V_I E_R E_H(\hat{\theta}|\mathbf{n}, \mathbf{m}) + E_I V_R E_H(\hat{\theta}|\mathbf{n}, \mathbf{m}) + E_I E_R V_H(\hat{\theta}|\mathbf{n}, \mathbf{m}).$$

Proof: Proof for the expectation and variance of the estimator for post-stratified random sampling is similar to that of lemma 5.1. ■

Lemma 5.4 *Expectation for Variance with Response in the Post-stratified Sample*

The expected variance with response in the post-stratified sample is the post-stratified sample variance for sampling with or without replacement:

$E(s_{hm}^2|\mathbf{n}, \mathbf{m}) = s^2$, where $s_{hm}^2 = \frac{1}{m_h - 1}\sum_{k\in R_h}(y_{hk} - \hat{\mu}_{hm})^2$, $s_h^2 = \frac{1}{n_h - 1}\sum_{k=1}^{n}(y_{hk} - \hat{\mu}_h)^2$, $\hat{\mu}_{hm} = \frac{1}{m_h}\sum_{k=1}^{m_h} y_{hk}$ where R_h is the response indicator in the sample and m is the sample and response size.

Proof: Proof is similar to that for lemma 5.2. This proof is assumed with MAR.

■

In this chapter for simplicity and convenience let us label the first n sampling units, $k = 1, ..., n$ as sampled, and the first $m < n$ sampling units as respondents.

5.2.1 Random Imputation

In this section, I present random hot-deck imputation in two subsections. Section 5.2.1.1 present theorems for the naive model. The *RHG* models are presented in section 5.2.1.2.

5.2.1.1 Naive Models with Random Imputation

A simple random sample of size n is selected from a population size N and m out of the n sample units respond. Nonrespondent samples are randomly replaced by values from responding sample units. Under the naive model with random imputation method, *Little & Rubin* (1987) give the proof of theorems 5.1 and 5.2. I extend the idea and prove for the basic sampling design in theorems 5.3-5.8.

To obtain the estimated variance for equal probability sampling design in theorems 5.1 and 5.2 replace σ^2 for sampling with replacement or S^2 for sampling without replacement with $s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k - \hat{\mu}_m)^2$, where $\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} y_k$ and m is the response size in the sample. The estimated variances are similarly found in theorems 5.3 and 5.4. For remaining theorems, estimates of variance are made as in theorem 2.7 or 2.8.

Theorem 5.1 *Simple random sampling with replacement under the naive model with random imputation*

In simple random sampling with replacement with random imputation under the naive model, an unbiased estimator of the mean μ is

$$\hat{\mu}_{ran,na}^{srs} = \frac{m\hat{\mu}_m + (n-m)\hat{\mu}_r}{n}, \tag{5.1}$$

where $\hat{\mu}_m$ and $\hat{\mu}_r$ are the mean of the responding units and of the imputed values respectively, with a variance of

$$V(\hat{\mu}_{ran,na}^{srs}) = \sigma^2[(\frac{1}{n} + \frac{1}{m}) + (\frac{m-1}{m})(\frac{n-m}{n^2})]. \tag{5.2}$$

Proof: The mean estimator in equation 5.1 is written as:

$$\hat{\mu}^{srs}_{ran,na} = \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} H_k y_k],$$

where H_k is the number of times y_k is used as a substitute for a missing value Y that $\sum_{k=1}^{m} H_k = n-m$, the number of nonrespondents. Conditioning on the sampled and responded values, the distribution of $(H_1, ..., H_m)$ in repeated random imputation is multinomial with sample size $n-m$ and probabilities (1/m,...,1/m) (See *Cochran*, 1977, section 2.10). Thus,

$$\begin{aligned} E(H_k|Y, X, n, m) &= (\frac{n-m}{m}), \\ V(H_k|Y, X, n, m) &= (n-m)(\frac{1}{m})(1-\frac{1}{m}), \end{aligned}$$

and for $k \neq i$,

$$COV(H_k H_i|Y, X, n, m) = -\frac{(n-m)}{m^2}.$$

By lemmas 4.2, 4.3, 5.1 and theorem 4.1,

$$\begin{aligned} E(\hat{\mu}^{srs}_{ran,na}) &= E_I E_R E_H \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} H_k y_k] \\ &= E_I E_R \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} (\frac{n-m}{m}) y_k] \\ &= E_I E_R \frac{1}{m} \sum_{k=1}^{m} y_k \\ &= E_I E_R \hat{\mu}_m \\ &= E_I \hat{\mu} \\ &= \mu. \end{aligned}$$

To prove equation 5.2 note that

$$\begin{aligned} V_I E_R E_H (\hat{\mu}^{srs}_{ran,na}) &= V_I E_R E_H \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} H_k y_k] \\ &= V_I(\hat{\mu}) \\ &= \frac{\sigma^2}{n}, \end{aligned}$$

$$\begin{aligned} E_I V_R E_H(\hat{\mu}^{srs}_{ran,na}) &= E_I V_R E_H \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} H_k y_k] \\ &= E_I V_R(\hat{\mu}_m) \\ &= E_I(\frac{s^2}{m}) \\ &= \frac{\sigma^2}{m}, \end{aligned}$$

and by lemma 5.2,

$$\begin{aligned} E_I E_R V_H(\hat{\mu}^{srs}_{ran,na}) &= E_I E_R V_H \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{m} H_k y_k] \\ &= E_I E_R \frac{1}{n^2} V_H[\sum_{k=1}^{m} H_k y_k] \\ &= E_I E_R \frac{1}{n^2}[\sum_{k=1}^{m} V_H(H_k) y_k^2 + \sum_{k=1}^{m}\sum_{i\neq k}^{m} COV(H_k H_i) y_k y_i] \\ &= \frac{1}{n^2} E_I E_R[\sum_{k=1}^{m}(n-m)\frac{1}{m}(1-\frac{1}{m}) y_k^2 - \sum_{k=1}^{m}\sum_{i\neq k}^{m}\frac{n-m}{m^2} y_k y_i] \\ &= \frac{1}{n^2}(\frac{n-m}{m^2}) E_I E_R[\sum_{k=1}^{m}(m-1) y_k^2 - \sum_{k=1}^{m}\sum_{i\neq k}^{m} y_k y_i] \\ &= \frac{1}{n^2}(\frac{n-m}{m^2}) E_I E_R[m\sum_{k=1}^{m} y_k^2 - (\sum_{k=1}^{m} y_k)^2] \\ &= \frac{1}{n^2}(\frac{n-m}{m}) E_I E_R[\sum_{k=1}^{m} y_k^2 - \frac{(\sum_{k=1}^{m} y_k)^2}{m}] \\ &= \frac{1}{n^2}(\frac{n-m}{m}) E_I E_R[(m-1) s_m^2] \\ &= \frac{1}{n^2}(\frac{n-m}{m})(m-1) E_I(s^2) \\ &= \frac{1}{n^2}(\frac{n-m}{m})(m-1)\sigma^2. \end{aligned}$$

Equation 5.2 follows by lemma 5.1. ∎

In sampling without replacement scheme, random imputation is applicable when nonrespondents is less than respondents, $n-m < m$. This shows that if nonresponse rate is higher than 50 % random imputation method cannot be used. Thus random imputed values are selected without replacement with notation $H_k = 1$ or 0 according to whether unit k is selected.

Theorem 5.2 *Simple random sampling without replacement under the naive model with random imputation*

In simple random sampling without replacement under the naive model with random imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}_{ran,na}^{srs} = \frac{(k^{'}+1)m\hat{\mu}_m + t\hat{\mu}_t}{n}, \tag{5.3}$$

where $n - m = k^{'}m + t$, $k^{'}$ is the number of times that all the response units are selected and t is the number of additional units selected to yield the $n - m$ nonresponse units and $0 \leq t < m$, $\hat{\mu}_m$ and $\hat{\mu}_t$ are the mean of responding units and of the t supplementary values of Y, with a variance of

$$V(\hat{\mu}_{ran,na}^{srs}) = (\frac{1}{m} - \frac{1}{N})S^2 + \frac{t}{n}(1 - \frac{t}{m})\frac{S^2}{n}. \tag{5.4}$$

Proof: By the theory of simple random sampling,

$$E_H(\hat{\mu}_t|Y, X, n, m) = \hat{\mu}_m \text{ and } V_H(\hat{\mu}_t|Y, X, n, m) = (1 - \tfrac{t}{m})\tfrac{s_m^2}{t},$$

where $\hat{\mu}_t = \frac{1}{t}\sum_{k=1}^{t} y_k = \frac{1}{t}\sum_{k=1}^{m} H_k y_k$. By lemmas 4.2, 4.3, 5.1 and theorem 4.2,

$$\begin{aligned} E(\hat{\mu}_{ran,na}^{srs}) &= E_I E_R E_H[\frac{(k^{'}+1)m\hat{\mu}_m + t\hat{\mu}_t}{n}] \\ &= E_I E_R[\frac{(k^{'}+1)m\hat{\mu}_m + t\hat{\mu}_m}{n}] \\ &= E_I E_R \hat{\mu}_m \\ &= E_I \hat{\mu} \\ &= \mu, \end{aligned}$$

where $n - m = k^{'}m + t$.

To prove equation 5.4 note that

$$\begin{aligned} V_I E_R E_H(\hat{\mu}_{ran,na}^{srs}) &= V_I(\hat{\mu}) \\ &= (\frac{1}{n} - \frac{1}{N})S^2, \end{aligned}$$

$$\begin{aligned} E_I V_R E_H(\hat{\mu}^{srs}_{ran,na}) &= E_I V_R(\hat{\mu}_m) \\ &= E_I(\frac{1}{m} - \frac{1}{n})s^2 \\ &= (\frac{1}{m} - \frac{1}{n})S^2, \end{aligned}$$

and by lemma 5.2,

$$\begin{aligned} E_I E_R V_H(\hat{\mu}^{srs}_{ran,na}) &= E_I E_R(\frac{t}{n})^2 V_H(\hat{\mu}_t) \\ &= E_I E_R \frac{t^2}{n^2}(\frac{m-t}{m})\frac{s^2_m}{t} \\ &= E_I(\frac{t^2}{n^2})(1 - \frac{t}{m})\frac{s^2}{t} \\ &= (\frac{t^2}{n^2})(1 - \frac{t}{m})\frac{S^2}{t}. \end{aligned}$$

Equation 5.4 follows by lemma 5.1.

■

Theorem 5.3 *Stratified random sampling with replacement under the naive model with random imputation*

In stratified random sampling with replacement under the naive model with random imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{st}_{ran,na} = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})\frac{1}{n_{hl}}[m_{hl}\hat{\mu}_{hlm} + (n_{hl} - m_{hl})\hat{\mu}_{hlr}], \tag{5.5}$$

where $\hat{\mu}_{hlm}$ and $\hat{\mu}_{hlr}$ are the mean of the responding units and of the imputed values respectively, with a variance of

$$V(\hat{\mu}^{st}_{ran,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2\sigma^2_{hl}[(\frac{1}{n_{hl}} + \frac{1}{m_{hl}}) + (\frac{m_{hl}-1}{m_{hl}})(\frac{n_{hl}-m_{hl}}{n^2_{hl}})]. \tag{5.6}$$

Proof: Sampling in one stratum is independent of sampling in another stratum, so that the stratum mean estimators $\hat{\mu}^{srs}_{hl,ran,na}$ are mutually independent. By theorem 5.1, an unbiased estimator of the population mean and its variance in equation 5.5 and 5.6 are

$$E(\hat{\mu}^{st}_{ran,na} = sum^{H}_{h=1} \sum^{L}_{l=1} \frac{N_{hl}}{N} E\hat{\mu}^{srs}_{hl,ran,na}$$

and

$$V(\hat{\mu}^{st}_{ran,na} = sum^{H}_{h=1} \sum^{L}_{l=1} (\frac{N_{hl}}{N})^2 V\hat{\mu}^{srs}_{hl,ran,na}$$

■

Theorem 5.4 *Stratified random sampling without replacement under the naive model with random imputation*

In stratified random sampling without replacement under the naive model with random imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{st}_{ran,na} = \sum_{h=1}^{H} \sum_{l=1}^{L} (\frac{N_{hl}}{N}) \frac{1}{n_{hl}} [m_{hl}\hat{\mu}_{hlm} + (n_{hl} - m_{hl})\hat{\mu}_{hlt}], \tag{5.7}$$

where $n_{hl} - m_{hl} = k_{hl}m_{hl} + t_{hl}$, k_{hl} is the number of times that all the response units in stratum hl are selected and t_{hl} is the number of additional units selected to yield the $n_{hl} - m_{hl}$ nonresponse units and $0 \leq t_{hl} < m_{hl}$, $\hat{\mu}_{hlm}$ and $\hat{\mu}_{hlt}$ are the mean of the responding units and of the t supplementary values of Y in stratum hl respectively, with a variance of

$$V(\hat{\mu}^{st}_{ran,na}) = \sum_{h=1}^{H} \sum_{l=1}^{L} (\frac{N_{hl}}{N})^2 [(\frac{1}{m_{hl}} - \frac{1}{N_{hl}}) S^2_{hl} + (\frac{t_{hl}}{n_{hl}})(1 - \frac{t_{hl}}{m_{hl}}) \frac{S^2_{hl}}{n_{hl}}]. \tag{5.8}$$

Proof: Proof for unbiaseness of the mean estimator and for its variance is similar to that of theorem 5.3 by using theorem 5.2.

■

Theorem 5.5 *Random Equal Probability Sampling with Replacement under the naive model with random imputation*

In random sampling with varying probability with replacement under the naive model and random imputation, an unbiased estimator of the total τ is

$$\hat{\tau}^{srs,pps}_{ran,na} = \frac{m\tilde{\tau}_m + (n-m)\tilde{\tau}_r}{n}, \tag{5.9}$$

where $\tilde{\tau}_m$ and $\tilde{\tau}_r$ are the mean of the responding units and of the imputed values $\tilde{y}_k = \frac{y_k}{p_k}$ respectively, with a variance of

$$V(\hat{\tau}_{ran,na}^{srs,pps}) = \frac{1}{n}[\sum_{k=1}^{N} \frac{y_k^2}{p_k} - \tau^2] + E_I \frac{s_{\tilde{y}_m}^2}{m} + E_I E_R \frac{n-m}{n^2 m^2}(m-1)s_{\tilde{y}_m}^2, \tag{5.10}$$

where τ is the total of variable of interests, $s_{\tilde{y}_m}^2$ is a variance of $\tilde{y}_k$.

Proof: Proof for unbiaseness of the total estimator and for its variance is similar to that for theorem 5.1 by replacing y_k with $\frac{y_k}{p_k}$ and using the results of theorem 4.8.

■

Theorem 5.6 *Random Equal Probability Sampling without Replacement under the naive model with random imputation*

In random sampling with varying probability without replacement under the naive model and random imputation, an unbiased estimator of the total τ is

$$\tau_{ran,na}^{srs,\pi ps} = (k^{'} + 1)m\tilde{\tau}_m + t\tilde{\tau}_t, \tag{5.11}$$

$k^{'}$ is the number of times that all the response units are selected and t is the number of additional units selected to yield the $n-m$ nonresponse units and $0 \leq t < m$, $\tilde{\tau}_m = \frac{1}{m}\sum_{k=1}^{m} \frac{y_k}{\pi_k}$ and $\tilde{\tau}_t = \frac{1}{t}\sum_{k=1}^{t} \frac{y_k}{\pi_k}$, $n - m = k^{'}m + t$, with a variance of

$$V(\hat{\tau}_{ran,na}^{srs,\pi ps}) = \sum_{k=1}^{N}(\frac{1-\pi_k}{\pi_k})y_k^2 + \sum_{k=1}^{N}\sum_{i\neq k}^{N}(\frac{\pi_{ki}-\pi_k\pi_i}{\pi_k\pi_i})y_k y_i +$$

$$E_I n^2(1-f_1)\frac{s_{\tilde{y}_m}^2}{m} + E_I E_R(1-f_2)ts_{\tilde{y}_t}^2, \tag{5.12}$$

where $f_1 = \frac{m}{n}$, $f_2 = \frac{t}{m}$, $s_{\tilde{y}_m}^2 = \frac{1}{m-1}\sum_{k=1}^{m}(\tilde{y}_k - \tilde{\tau}_m)^2$ and $s_{\tilde{y}_t}^2 = \frac{1}{t-1}\sum_{k=1}^{t}(\tilde{y}_k - \tilde{\tau}_t)^2$.

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that theorem 5.2 by replacing y_k with $\frac{y_k}{\pi_k}$ and using the results of theorem 4.7.

■

Theorem 5.7 *Stratified Equal Probability Sampling with Replacement under the naive model with random imputation*

In stratified random sampling with varying probability with replacement under the naive model with random imputation, an unbiased estimator of the total τ is

$$\hat{\tau}_{ran,na}^{st,pps} = \sum_{h=1}^{H}\sum_{l=1}^{L}\{\frac{m_{hl}\tilde{\tau}_{hlm} + (n_{hl} - m_{hl})\tilde{\tau}_{hlr}}{n_{hl}}\}, \tag{5.13}$$

where $\tilde{\tau}_{hlm}$ and $\tilde{\tau}_{hlr}$ are the mean of the responding units and of the imputed values $\tilde{y}_{hlk} = \frac{y_{hlk}}{p_{hlk}}$ in stratum hl respectively, with a variance of

$$V(\hat{\tau}_{ran,na}^{st,pps}) = \sum_{h=1}^{H}\sum_{l=1}^{L}\{\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}^2}{p_{hlk}} - \tau_{hl}^2] + E_I\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}} + E_I E_R\frac{n_{hl} - m_{hl}}{n_{hl}^2 m_{hl}^2}(m_{hl} - 1)s_{\tilde{y}_{hlm}}^2\}, \tag{5.14}$$

where τ_{hl} and $s_{\tilde{y}_{hlm}}^2$ are the total of variable of interests and the variance of $\tilde{y}_{hlk}$ in stratum hl respectively.

Proof: Proof for unbiasedness of the total estimator and its variance is similar to that for theorem 5.3.

■

Theorem 5.8 *Stratified Equal Probability Sampling without Replacement under the naive model with random imputation*

In stratified random sampling with varying probability without replacement under the naive model with random imputation, an unbiased estimator of the total τ is

$$\tau_{ran,na}^{st,\pi ps} = \sum_{h=1}^{H}\sum_{l=1}^{L}\{(k_{hl}^{'} + 1)m_{hl}\tilde{\tau}_{hlm} + t\tilde{\tau}_{hlt}\}, \tag{5.15}$$

where $k_{hl}^{'}$ and t_{hl} are the number of times that all the response units are selected and the number of additional units selected to yield the $n_{hl} - m_{hl}$ nonresponse units and $0 \leq t_{hl} < m_{hl}$ in stratum hl respectively, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}}$, $\tilde{\tau}_{hlt} = \frac{1}{t_{hl}}\sum_{k=1}^{t_{hl}}\frac{y_{hlk}}{\pi_{hlk}}$ and $n_{hl} - m_{hl} = k_{hl}^{'}m_{hl} + t_{hl}$, with a variance of

$$V(\hat{\tau}_{ran,na}^{st,\pi ps}) = \sum_{h=1}^{H}\sum_{l=1}^{L}\{\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y_{hlk}^2 + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli} +$$

$$E_I n_{hl}^2(1-f_{hl1})\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}} + E_I E_R(1-f_{hl2})t_{hl}s_{\tilde{y}_{hlt}}^2\}, \tag{5.16}$$

where

$$f_{hl1} = \frac{m_{hl}}{n_{hl}},\ f_{hl2} = \frac{t_{hl}}{m_{hl}},\ s_{\tilde{y}_{hlm}}^2 = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk} - \tilde{\tau}_{m_{hl}})^2 \text{ and}$$

$$s_{\tilde{y}_{hlt}}^2 = \frac{1}{t_{hl}-1}\sum_{k=1}^{t_{hl}}(\tilde{y}_{hlk} - \tilde{\tau}_{hlt})^2.$$

Proof: Proof for unbiasedness of the total estimator and its variance is similar to that for theorem 5.3.

■

5.2.1.2 *RHG* Models in Random Imputation

I extend and prove random imputation with the *RHG* model in theorems 5.9-5.12. To obtain the estimated variance for equal probability sampling design in theorems 5.9-5.10 replace σ_{hl}^2 for sampling with replacement or S_{hl}^2 for sampling without replacement with $s_{hlm}^2 = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(y_{hlk} - \hat{\mu}_{hlm})^2$, where $\hat{\mu}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}} y_{hlk}$ and m_{hl} is the response size in the post-stratum *hl*. For remaining theorems, estimates of variance are made as in theorem 2.7 or 2.8.

Theorem 5.9 *Simple random sampling with replacement under the RHG model with random imputation*

In simple random sampling with replacement under the *RHG* model with random imputation, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}_{ran,RHG}^{srs} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})[\frac{m_{hl}\hat{\mu}_{hlm} + (n_{hl}-m_{hl})\hat{\mu}_{hlr}}{n_{hl}}], \tag{5.17}$$

where $\hat{\mu}_{hlm}$ and $\hat{\mu}_{hlr}$ are the mean of the responding units and of the imputed values in stratum hl respectively, with a conditional variance of

$$V(\hat{\mu}^{srs}_{ran,RHG}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\sigma^2_{hl}[(\frac{1}{n_{hl}}+\frac{1}{m_{hl}})+(\frac{m_{hl}-1}{m_{hl}})(\frac{n_{hl}-m_{hl}}{n^2_{hl}})]. \quad (5.18)$$

Proof: The sample is post-stratified into $H_S \times L_S$ post-strata. By theorem 5.1,

$$\begin{aligned} E(\hat{\mu}^{srs}_{ran,RHG}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}E(\hat{\mu}^{srs}_{hl,ran,na}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\mu_{hl} \\ &= \mu, \end{aligned}$$

and a conditional variance of mean estimator is

$$\begin{aligned} V(\hat{\mu}^{srs}_{ran,RHG}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2V(\hat{\mu}^{srs}_{hl,ran,na}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2[\sigma^2_{hl}[\frac{1}{n_{hl}}+\frac{1}{m_{hl}}+(\frac{m_{hl}-1}{m_{hl}})(\frac{n_{hl}-m_{hl}}{n^2_{hl}})]]. \end{aligned}$$

■

Theorem 5.10 *Simple random sampling without replacement under RHG model with random imputation*

In simple random sampling without replacement under *RHG* model with random imputation, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}^{srs}_{ran,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})\frac{(k'_{hl}+1)m_{hl}\hat{\mu}_{hlm}+t_{hl}\hat{\mu}_{hlt}}{n_{hl}}, \quad (5.19)$$

where $n_{hl}-m_{hl} = k'_{hl}m_{hl}+t_{hl}$, k'_{hl} is the number of times that all the response units in stratum hl are selected and t_{hl} is the number of additional units selected to yield the $n_{hl}-m_{hl}$ nonresponse units and $0 \le t_{hl} < m_{hl}$, $\hat{\mu}_{hlm}$ and $\hat{\mu}_{hlt}$ are the mean of the responding units and of the t supplementary values of Y in stratum hl respectively, with a conditional variance of

$$V(\hat{\mu}^{srs}_{ran,RHG}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2[(\frac{1}{m_{hl}}-\frac{1}{N_{hl}})S^2_{hl}+\frac{t^2_{hl}}{n^2_{hl}}(1-\frac{t_{hl}}{m_{hl}})\frac{S^2_{hl}}{t_{hl}}]. \quad (5.20)$$

Proof: Proof for unbiasedness of the mean estimator and its variance is similar to that for theorem 5.9.

■

Theorem 5.11 *Random Equal Probability with Replacement under the RHG model and random imputation*

In random sampling with varying probability with replacement under the *RHG* model and random imputation, a conditional unbiased estimator of the total τ is

$$\hat{\tau}^{srs,pps}_{ran,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\{\frac{m_{hl}\tilde{\tau}_{hlm} + (n_{hl} - m_{hl})\tilde{\tau}_{hlr}}{n_{hl}}\}, \tag{5.21}$$

where $\tilde{\tau}_{hlm}$ and $\tilde{\tau}_{hlr}$ are the mean of the responding units and of the imputed values $\tilde{y}_{hlk} = \frac{y_{hlk}}{p_{hlk}}$ in stratum hl respectively, with a conditional variance of

$$V(\hat{\tau}^{srs,pps}_{ran,RHG}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\{\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y^2_{hlk}}{p_{hlk}} - \tau^2_{hl}] + E_I\frac{s^2_{\tilde{y}_{hlm}}}{m_{hl}} + E_I E_R\frac{n_{hl}-m_{hl}}{n^2_{hl}m^2_{hl}}(m_{hl}-1)s^2_{\tilde{y}_{hlm}}\}, \tag{5.22}$$

where τ_{hl} and $s^2_{\tilde{y}_{hlm}}$ are the total of variable of interests and the variance of $\tilde{y}_{hlk}$ in stratum hl respectively.

Proof: The sample is post-stratified into $H_S \times L_S$ post-strata. By theorem 5.5,

$$\begin{aligned} E(\hat{\tau}^{srs,pps}_{ran,RHG}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S} E(\hat{\tau}^{srs,pps}_{hl,ran,na}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=1}^{N_{hl}} y_{hlk} \\ &= \tau, \end{aligned}$$

and the conditional variance of the total estimator is

$$V(\hat{\tau}^{srs,pps}_{ran,RHG}|\mathbf{n},\mathbf{m}) = \textstyle\sum_{h=1}^{H_S}\sum_{l=1}^{L_S} V(\hat{\tau}^{srs,pps}_{hl,ran,na}).$$

■

Theorem 5.12 *Random Equal Probability Sampling without Replacement under the RHG model with random imputation*

In random sampling with varying probability without replacement under the *RHG* model and random imputation, a conditional unbiased estimator of the total τ is

$$\tau_{ran,RHG}^{srs,\pi ps} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\{(k_{hl}^{'}+1)m_{hl}\tilde{\tau}_{hlm} + t\tilde{\tau}_{hlt}\}, \tag{5.23}$$

where $k_{hl}^{'}$ and t_{hl} are the number of times that all the response units are selected and the number of additional units selected to yield the $n_{hl}-m_{hl}$ nonresponse units and $0 \leq t_{hl} < m_{hl}$ in stratum hl respectively, $\tilde{\tau}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}}$, $\tilde{\tau}_{hlt} = \frac{1}{t_{hl}}\sum_{k=1}^{t_{hl}}\frac{y_{hlk}}{\pi_{hlk}}$ and $n_{hl} - m_{hl} = k_{hl}^{'}m_{hl} + t_{hl}$, with a conditional variance of

$$V(\hat{\tau}_{ran,RHG}^{srs,\pi ps}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\{\sum_{k=1}^{N_{hl}}(\frac{1-\pi_{hlk}}{\pi_{hlk}})y_{hlk}^2 + \sum_{k=1}^{N_{hl}}\sum_{i\neq k}^{N}(\frac{\pi_{hlki}-\pi_{hlk}\pi_{hli}}{\pi_{hlk}\pi_{hli}})y_{hlk}y_{hli} +$$

$$E_I n_{hl}^2(1-f_{hl1})\frac{s_{\tilde{y}_{hlm}}^2}{m_{hl}} + E_I E_R(1-f_{hl2})t_{hl}s_{\tilde{y}_{hlt}}^2\}, \tag{5.24}$$

where

$$f_{hl1} = \tfrac{m_{hl}}{n_{hl}},\ f_{hl2} = \tfrac{t_{hl}}{m_{hl}},\ s_{\tilde{y}_{hlm}}^2 = \tfrac{1}{m_{hl}-1}\textstyle\sum_{k=1}^{m_{hl}}(\tilde{y}_{hlk} - \tilde{\tau}_{m_{hl}})^2 \text{ and}$$

$$s_{\tilde{y}_{hlt}}^2 = \tfrac{1}{t_{hl}-1}\textstyle\sum_{k=1}^{t_{hl}}(\tilde{y}_{hlk} - \tilde{\tau}_{hlt})^2.$$

Proof: Proof for unbiasedness of the total estimator and its variance is similar to that for theorem 5.11.

■

5.2.2 Sequential Imputation

In this section, I present sequential hot-deck imputation in two subsections. Section 5.2.2.1 presents theorems for the naive model. The *RHG* models are presented in section 5.2.2.2.

To obtain the estimated variance for equal probability sampling design in theorems 5.13 replace σ^2 with $s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k - \hat{\mu}_m)^2$, where $\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} y_k$ and m is the response size in the sample. The estimated variances are similarly found in theorems 5.14 and 5.15.

5.2.2.1 Naive Models with Sequential Imputation

A simple random sample of size n is selected from a population size N and m out of the n sample units respond. Response and nonresponse units are treated in a sequence, and a missing data Y is replaced by the nearest responding value preceding it in the sequence. If the first sample unit is nonresponse, then the value called the *cold-deck* value, y_0, is taken from the previous survey where the auxiliary information is similar to that of the nonrespondent. Under the naive model with sequential imputation method, *Bailar et al* (1978) give the proof of theorem 5.13 in the article *"A comparison of some adjustment and weighting procedure for survey data"*. I state theorem 5.13 without proof. I extend the idea and prove for stratified random sampling in theorem 5.14.

Theorem 5.13 *Simple random sampling with replacement under the naive model and sequential imputation*

Let y_0 be the cold-deck value obtained from previous surveys and let c_k be the number of times the k^{th} value is used with $\sum_{k=0}^{n} c_k = n$. Assume y and c are independent for all observed units. Then with simple random sampling with replacement under the naive model and sequential imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}_{seq,na}^{srs} = \frac{1}{n}\sum_{k=0}^{n} c_k y_k, \tag{5.25}$$

with a variance of

$$V(\hat{\mu}_{seq,na}^{srs}) = \frac{\sigma^2}{n}[1 + \{\frac{2(n-m)}{n}\}\{\frac{(mn+n-1)}{(m+1)(m+2)}\}]. \tag{5.26}$$

■

Theorem 5.14 *Stratified random sampling with replacement under the naive model and sequential imputation*

Let y_{hl0} be the cold-deck value obtained from previous surveys and let c_{hlk} be the number of times the k^{th} value in stratum hl is used with $\sum_{h=1}^{H}\sum_{l=1}^{L}\sum_{k=0}^{n_{hl}} c_{hlk} = n$. Assume y_{hl} and c_{hl} are independent for all observed units in stratum hl. Then with stratified random sampling with replacement under the naive model and sequential imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{st}_{seq,na} = \sum_{h=1}^{H}\sum_{l=1}^{L}\frac{N_{hl}}{Nn_{hl}}\sum_{k=0}^{n_{hl}} c_{hlk}y_{hlk}, \tag{5.27}$$

with a variance of

$$V(\hat{\mu}^{st}_{seq,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2\frac{\sigma^2_{hl}}{n_{hl}}[1+\{\frac{2(n_{hl}-m_{hl})}{n_{hl}}\}\{\frac{(m_{hl}n_{hl}+n_{hl}-1)}{(m_{hl}+1)(m_{hl}+2)}\}]. \tag{5.28}$$

Proof: Sampling in one stratum is independent of sampling in another stratum, so that the stratum mean estimator $\hat{\mu}^{srs}_{hl,seq,na}$ are mutually independent. By thoerem 5.13, an unbiased estimator of the population mean and its variance in equation 5.27 and 5.28 are

$$E(\hat{\mu}^{st}_{seq,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})E[\hat{\mu}^{srs}_{hl,seq,na}]$$

and

$$V(\hat{\mu}^{st}_{seq,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2V[\hat{\mu}^{srs}_{hl,seq,na}]$$

■

5.2.2.2 *RHG* Model in Sequential Imputation

I extend and prove sequential imputation with the *RHG* model in theorem 5.15.

Theorem 5.15 *Simple random sampling with replacement under the RHG model and sequential imputation*

Let y_{hl0} be the cold-deck value that can get from the previous survey and let c_{hlk} be the number of times the k^{th} value in stratum hl is used with $\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\sum_{k=0}^{n_{hl}} c_{hlk} =$

n. Assume y_{hl} and c_{hl} are independent for all observed units in stratum hl. Then with simple random sampling with replacement under the *RHG* model and sequential imputation, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}^{srs}_{seq,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nn_{hl}}\sum_{k=0}^{n_{hl}} c_{hlk}y_{hlk}, \tag{5.29}$$

with a conditional variance of

$$V(\hat{\mu}^{srs}_{seq,RHG}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{\sigma^2_{hl}}{n_{hl}}[1+\{\frac{2(n_{hl}-m_{hl})}{n_{hl}}\}\{\frac{(m_{hl}n_{hl}+n_{hl}-1)}{(m_{hl}+1)(m_{hl}+2)}\}]. \tag{5.30}$$

Proof: The sample is post-stratified into $H_S \times L_S$ post-strata. By theorem 5.13,

$$\begin{aligned} E(\hat{\mu}^{srs}_{seq,RHG}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}E(\hat{\mu}^{srs}_{hl,seq,na}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\mu_{hl} \\ &= \mu. \end{aligned}$$

The conditional variance of $\hat{\mu}^{srs}_{seq,RHG}$ given $\mathbf{n},\mathbf{m}$ is

$$\begin{aligned} V(\hat{\mu}^{srs}_{seq,RHG}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V(\hat{\mu}^{srs}_{hl,seq,na}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{\sigma^2_{hl}}{n_{hl}}[1+\{\frac{2(n_{hl}-m_{hl})}{n_{hl}}\}\{\frac{(m_{hl}n_{hl}+n_{hl}-1)}{(m_{hl}+1)(m_{hl}+2)}\}]. \end{aligned}$$

■

5.2.3 Stochastic Regression Imputation

Kalton & Kasprzyk (1982) suggest that most explicit imputation methods can be expressed through a model linking the auxiliary data to the value of a missing item. If Z is the imputed value and $\mathbf{X} = (X_{1_i},...,X_{k_i})$ is the $k-$dimensional vector of continuous or categorical auxiliary variables for the i^{th} nonrespondent with actual Y_i, the general imputation model,

$$Z_i = f(\mathbf{X}) + e_i,$$

can be used to describe most explicit methods, where $f(.)$ is some function of the auxiliary data and e_i is a specified residual [the stochastic part of the term stochastic regression imputation]. The function form of $f(.)$ is almost always linear, so that this framework may be expressed the i^{th} imputed nonrespondent as

$$Z_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ji} + e_i,$$

where the $\hat{\beta}$ is ordinary least square estimators.

If $e_i = 0$ is specified, then the case Z_i is a deterministic prediction given the respondent data. This may distort the shape of the distribution of the Y variable or inflate the degree of association between the Y variable and the set of assignment variables, $\mathbf{X}$. To solve this problem, random residuals, $e_i \neq 0$, are designed with zero expectation.

Kalton (1983) lists several ways in which a random residual, e_i, can be identified. A first way is to take all residuals from the same distribution with zero mean and residual variance as estimated in fitting the assumed model to respondent data, e.g., normal distribution. A second way is to choose the residual by applying the first approach separately within certain subpopulation. A third alternative way is to use the residual of the fitted model from a randomly chosen respondent and the fourth way is the same as the third except that the donor residual is chosen from respondents with similar values of the assignment variables.

Thus, stochastic regression imputation replaces missing values by a value predicted by regression imputation plus a residual which is drawn to reflect uncertainty in the predicted value.

Schaible (1983) shows an application of regression imputation with continuous Y variable in simple random sampling and investigates the properties of deterministic

regression imputation. *Herzog & Rubin* (1983) describe a two-stage procedure for regression imputation to missing Social Security income data in survey of low-income aged and disabled.

Greenless et al (1982) and *David et al* (1986) illustrate the use of regression imputation for dealing with earnings income missing from respondents to the Current Population Survey.

In this section stochastic simple linear regression with the residual of the fitted model from a randomly chosen respondent is used to compensate nonrespondents. I present stochastic regression imputation in two subsections. Section 5.2.3.1 presents theorems for the naive model. The *RHG* models are presented in section 5.2.3.2.

5.2.3.1 Naive Model in Stochastic Regression Imputation

In simple random sampling with or without replacement scheme, a sample size n is selected from a population size N and m out of the n sampled units respond. Nonrespondent samples are replaced by the imputed value from stochastic simple linear regression. *Schaible* (1983) gives the result of deterministic regression imputation. I extend and prove to stochastic regression imputation in theorems 5.16-5.23.

To obtain the estimated variance for equal probability sampling design in theorems 5.16 and 5.17 replace σ^2 for sampling with replacement or S^2 for sampling without replacement with $s_m^2 = \frac{1}{m-1}\sum_{k=1}^{m}(y_k - \hat{\mu}_m)^2$, where $\hat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} y_k$ and m is the response size in the sample. The estimated variances are similarly found in theorems 5.18 and 5.19. For remaining theorems, estimates of variance are made as in theorem 2.7 or 2.8.

Theorem 5.16 *Simple random sampling with replacement under the naive model and stochastic regression imputation*

In simple random sampling with replacement under the naive model and stochastic regression imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}_{reg,na}^{srs} = \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{r} z_k], \tag{5.31}$$

where $z_k = \hat{\beta}_0 + \hat{\beta}_k x_k + e_k$, $r = n - m$, $\hat{\beta}_0 = \hat{\mu}_{ym} - \hat{\beta}_1 \hat{\mu}_{xm}$, and $\hat{\beta}_1 = \frac{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2(y_k - \hat{\mu}_{ym})^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2}$, with a variance of

$$V(\hat{\mu}^{srs}_{reg,na}) = \frac{\sigma^2}{n} + (\frac{r\sigma}{n})^2[\frac{(\hat{\mu}_{xm} - \hat{\mu}_{xr})^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} + \frac{1}{m} + \frac{1}{r}], \tag{5.32}$$

where $\hat{\mu}_{xm}$ and $\hat{\mu}_{xr}$ are the response and nonresponse sample mean of the auxiliary variable X respectively and $\hat{\mu}_{ym}$ is the response sample mean of the study variable Y.

Proof: The estimator $\hat{\mu}$ is a function of random variable Y, X and Z, indicator I and R. $\hat{\mu}$ can be written as

$$\hat{\mu} = \tfrac{1}{n}[\textstyle\sum_{k=1}^{N} I_k R_k y_k] + \sum_{k=1}^{N} I_k(1 - R_k) z_k,$$

where $z_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + e_k$.

Note that the properties of e_k are *iid* $N(0, \sigma^2)$ and independent of X_k and Y_k. $\hat{\beta}_0$, $\hat{\beta}_1$ are unbiased estimator of β_0 and β_1. Thus the expected of $\hat{\mu}$ is

$$\begin{aligned}
E(\hat{\mu}^{srs}_{reg,na}) &= E_I E_R E_Z(\hat{\mu}|Y, X, Z, I, R) \\
&= E_I E_R E_Z \frac{1}{n}[\sum_{k=1}^{N} I_k R_k y_k + \sum_{k=1}^{N} I_k(1 - R_k) z_k] \\
&= \frac{1}{n} E_I E_R[\sum_{k=1}^{N} I_k R_k y_k + \sum_{k=1}^{N} I_k(1 - R_k) E_Z(\hat{\beta}_0 + \hat{\beta}_1 x_k + e_k)] \\
&= \frac{1}{n} E_I E_R[\sum_{k=1}^{N} I_k R_k y_k + \sum_{k=1}^{N} I_k(1 - R_k)(\beta_0 + \beta_1 x_k)] \\
&= \frac{1}{n} E_I E_R[\sum_{k=1}^{N} I_k R_k y_k + \sum_{k=1}^{N} I_k(1 - R_k) y_k] \\
&= \frac{1}{n} E_I E_R[\sum_{k=1}^{N} I_k y_k] \\
&= \frac{1}{n} \sum_{k=1}^{N} \frac{n}{N} y_k \\
&= \mu.
\end{aligned}$$

The variance of $\hat{\mu}$ can be expressed as

$$V(\hat{\mu}) =$$

$$V_I E_R E_Z(\hat{\mu}|Y,X,Z,I,R) + E_I V_R E_Z(\hat{\mu}|Y,X,Z,I,R) + E_I E_R V_Z(\hat{\mu}|Y,X,Z,I,R).$$

To prove equation 5.32 note that

$$\begin{aligned} V_I E_R E_Z(\hat{\mu}^{srs}_{reg,na}) &= V_I(\hat{\mu}) \\ &= \frac{\sigma^2}{n}, \end{aligned}$$

$$\begin{aligned} E_I V_R E_Z(\hat{\mu}^{srs}_{reg,na}) &= \frac{1}{n} E_I V_R[\sum_{k=1}^{N} I_k Y_k] \\ &= 0, \end{aligned}$$

and $E_I E_R V_Z(\hat{\mu}^{srs}_{reg,na})$

$$\begin{aligned} &= E_I E_R V_Z \frac{1}{n}[\sum_{k=1}^{N} I_k R_k y_k + \sum_{k=1}^{N} I_k(1-R_k)z_k] \\ &= E_I E_R V_Z \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{r} z_k] \\ &= E_I E_R V_Z \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{r}(\hat{\beta}_0 + \hat{\beta}_1 x_k + e_k)] \\ &= E_I E_R (\frac{r}{n})^2[V_Z(\hat{\beta}_0) + \hat{\mu}^2_{xr} V_Z(\hat{\beta}_1) + V_Z(\hat{\mu}_e) + 2\hat{\mu}_{xr} COV(\hat{\beta}_0, \hat{\beta}_1)] \\ &= E_I E_R (\frac{r}{n})^2[\frac{\sigma^2 \sum_{k=1}^{m} x_k^2}{m \sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} + \frac{\hat{\mu}_r^2 \sigma^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} + \frac{\sigma^2}{r} - \frac{2\hat{\mu}_{xr}\hat{\mu}_{xm}\sigma^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2}] \\ &= (\frac{r\sigma}{n})^2[\frac{(\hat{\mu}_{xm} - \hat{\mu}_{xr})^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} + \frac{1}{m} + \frac{1}{r}]. \end{aligned}$$

Equation 5.32 follows variance expression.

■

Theorem 5.17 *Simple random sampling without replacement under the naive model and stochastic regression imputation*

In simple random sampling without replacement under the naive model and stochastic regression imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{srs}_{reg,na} = \frac{1}{n}[\sum_{k=1}^{m} y_k + \sum_{k=1}^{r} z_k], \tag{5.33}$$

where $z_k = \hat{\beta}_0 + \hat{\beta}_k x_k + e_k$, $r = n - m$, $\hat{\beta}_0 = \hat{\mu}_{ym} - \hat{\beta}_1 \hat{\mu}_{xm}$, and $\hat{\beta}_1 = \frac{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2(y_k - \hat{\mu}_{ym})^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2}$, with a variance of

$$V(\hat{\mu}^{srs}_{reg,na}) = \frac{N-n}{nN}S^2 + (\frac{rS}{n})^2[\frac{(\hat{\mu}_{xm} - \hat{\mu}_{xr})^2}{\sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} + \frac{1}{m} + \frac{1}{r}], \tag{5.34}$$

where $\hat{\mu}_{xm}$ and $\hat{\mu}_{xr}$ are the response and nonresponse sample mean of the auxiliary variable X respectively and $\hat{\mu}_{ym}$ is the response sample mean of the study variable Y.

Proof: Proof can be followed as theorem 5.16 except

$$V_I E_R E_Z(\hat{\mu}^{srs}_{reg,na}) = V_1(\hat{\mu}_y) = \tfrac{N-n}{nN}S^2.$$

■

Theorem 5.18 *Stratified random sampling with replacement under the naive model and stochastic regression imputation*

In stratified random sampling with replacement under the naive model and stochastic regression imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{st}_{reg,na} = \sum_{h=1}^{H}\sum_{l=1}^{L}\frac{N_{hl}}{Nn_{hl}}[\sum_{k=1}^{m_{hl}} y_{hlk} + \sum_{k=1}^{r_{hl}} z_{hlk}], \tag{5.35}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1}x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \hat{\mu}_{yhlm} - \hat{\beta}_{hl1}\hat{\mu}_{xhlm} \text{ and}$$
$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})(y_{hlk} - \hat{\mu}_{yhlm})}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2},$$

with a variance of

$$V(\hat{\mu}^{st}_{reg,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2[\frac{\sigma^2_{hl}}{n_{hl}} + (\frac{r_{hl}\sigma_{hl}}{n_{hl}})^2\{\frac{(\hat{\mu}_{xhlm} - \hat{\mu}_{xhlr})^2}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2} + \frac{1}{m_{hl}} + \frac{1}{r_{hl}}\}], \tag{5.36}$$

where $\hat{\mu}_{xhlm}$ and $\hat{\mu}_{xhlr}$ are the response and nonresponse sample mean of the auxiliary variable X in stratum hl respectively and $\hat{\mu}_{yhlm}$ is the response sample mean of the study variable Y in stratum hl.

Proof: Proof for unbiasedness of the mean estimator and for its variance is similar to that for theorem 5.3.

■

Theorem 5.19 *Stratified random sampling without replacement under the naive model and stochastic regression imputation*

In stratified random sampling without replacement under the naive model and stochastic regression imputation, an unbiased estimator of the mean μ is

$$\hat{\mu}^{st}_{reg,na} = \sum_{h=1}^{H}\sum_{l=1}^{L}\frac{N_{hl}}{Nn_{hl}}[\sum_{k=1}^{m_{hl}} y_{hlk} + \sum_{k=1}^{r_{hl}} z_{hlk}], \tag{5.37}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1}x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \hat{\mu}_{yhlm} - \hat{\beta}_{hl1}\hat{\mu}_{xhlm} \text{ and}$$
$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})(y_{hlk} - \hat{\mu}_{yhlm})}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2},$$

with a variance of

$$V(\hat{\mu}^{st}_{reg,na}) = \sum_{h=1}^{H}\sum_{l=1}^{L}(\frac{N_{hl}}{N})^2[\frac{N_{hl}-n_{hl}}{n_{hl}N_{hl}}S^2_{hl} + (\frac{r_{hl}S_{hl}}{n_{hl}})^2\{\frac{(\hat{\mu}_{xhlm} - \hat{\mu}_{xhlr})^2}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2} + \frac{1}{m_{hl}} + \frac{1}{r_{hl}}\}], \tag{5.38}$$

where $\hat{\mu}_{xhlm}$ and $\hat{\mu}_{xhlr}$ are the response and nonresponse sample mean of the auxiliary variable X in stratum hl respectively and $\hat{\mu}_{yhlm}$ is the response sample mean of the study variable Y in stratum hl.

Proof: Proof for unbiasedness of the mean estimator and for its variance is similar to that for theorem 5.3.

■

Theorem 5.20 *Random Equal Probability Sampling with with Replacement Selection under the naive model and stochastic regression imputation*

In random sampling with varying probability with replacement under the naive model and stochastic simple linear regression imputation, an unbiased estimator of the total τ is

$$\hat{\tau}^{srs,pps}_{reg,na} = \frac{1}{n}[\sum_{k=1}^{m}\frac{y_k}{p_k} + \sum_{k=1}^{r}\frac{z_k}{p_k}], \tag{5.39}$$

where

$$z_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + e_k,\ \hat{\beta}_0 = \frac{\sum_{k=1}^{m} \frac{y_k}{p_k} - \hat{\beta}_1 \sum_{k=1}^{m} \frac{x_k}{p_k}}{\sum_{k=1}^{m} \frac{1}{p_k}} \text{ and}$$

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{m} \frac{x_k y_k}{p_k} - (\sum_{k=1}^{m} \frac{x_k}{p_k} \sum_{k=1}^{m} \frac{y_k}{p_k}) / \sum_{k=1}^{m} \frac{1}{p_k}}{\sum_{k=1}^{m} \frac{x_k^2}{p_k} - (\sum_{k=1}^{m} \frac{x_k}{p_k})^2 / \sum_{k=1}^{m} \frac{1}{p_k}},$$

with a variance of

$$V(\hat{\tau}_{reg,na}^{srs,pps}) = \frac{1}{n}[\sum_{k=1}^{N} \frac{y_k}{p_k} - \tau^2] +$$

$$\frac{\sigma^2}{n^2 \sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2} \sum_{k=1}^{r} \frac{1}{p_k^2}[\frac{\sum_{i=1}^{m} x_i^2}{m} + x_k^2 - 2x_k\hat{\mu}_{xm} + \sum_{k=1}^{m}(x_k - \hat{\mu}_{xm})^2], \qquad (5.40)$$

where $\hat{\mu}_{xm}$ is the response sample mean of the auxiliary variable X.

Proof: The estimator $\hat{\tau}$ is a function of random variable Y, X and Z, indicator I and R. $\hat{\tau}$ can be written as

$$\hat{\tau} = \frac{1}{n}[\sum_{k=1}^{N} I_k R_k \frac{y_k}{p_k} + \sum_{k=1}^{N} I_k(1 - R_k)\frac{z_k}{p_k}],$$

where $z_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + e_k$.

Note that the properties of e_k are *iid* $N(0, \sigma^2)$ and independent of X_k and Y_k. $\hat{\beta}_0$, $\hat{\beta}_1$ are unbiased estimator of β_0 and β_1. Thus the expected of $\hat{\tau}$ is

$$\begin{aligned}
E(\hat{\tau}_{reg,na}^{srs}) &= E_I E_R E_Z(\hat{\tau}|Y, X, Z, I, R) \\
&= E_I E_R E_Z \frac{1}{n}[\sum_{k=1}^{N} I_k R_k \frac{y_k}{p_k} + \sum_{k=1}^{N} I_k(1 - R_k)\frac{z_k}{p_k}] \\
&= E_I E_R \frac{1}{n}[\sum_{k=1}^{N} I_k R_k \frac{y_k}{p_k} + \sum_{k=1}^{N} I_k(1 - R_k) E_Z \frac{(\hat{\beta}_0 + \hat{\beta}_1 x_k + e_k)}{p_k}] \\
&= E_I E_R \frac{1}{n}[\sum_{k=1}^{N} I_k R_k \frac{y_k}{p_k} + \sum_{k=1}^{N} I_k(1 - R_k)\frac{(\beta_0 + \beta_1 x_k)}{p_k}] \\
&= E_I E_R \frac{1}{n}[\sum_{k=1}^{N} I_k R_k \frac{y_k}{p_k} + \sum_{k=1}^{N} I_k(1 - R_k)\frac{y_k}{p_k}] \\
&= E_I E_R \frac{1}{n}[\sum_{k=1}^{N} I_k \frac{y_k}{p_k}] \\
&= \sum_{k=1}^{N} y_k \\
&= \tau.
\end{aligned}$$

The variance of $\hat{\tau}$ can be expressed as

$$V(\hat{\mu}) =$$

$$V_I E_R E_Z(\hat{\mu}|Y,X,Z,I,R) + E_I V_R E_Z(\hat{\mu}|Y,X,Z,I,R) + E_I E_R V_Z(\hat{\mu}|Y,X,Z,I,R).$$

To prove equation 5.40 note that

$$\begin{aligned} V_I E_R E_Z(\hat{\tau}_{reg,na}^{srs,pps}) &= V_I(\frac{1}{n}\sum_{k=1}^{n}\frac{y_k}{p_k}) \\ &= \frac{1}{n}[\sum_{k=1}^{N}\frac{y_k^2}{p_k} - \tau^2], \end{aligned}$$

$$\begin{aligned} E_I V_R E_Z(\hat{\tau}_{reg,na}^{srs,pps}) &= \frac{1}{n}[\sum_{k=1}^{N} I_k\frac{y_k}{p_k}] \\ &= 0, \end{aligned}$$

and $E_I E_R V_Z(\hat{\tau}_{reg,na}^{srs,pps})$

$$\begin{aligned} &= E_I E_R \frac{1}{n^2}\sum_{k=1}^{r}\frac{1}{p_k^2}V_Z(\hat{\beta}_0 + \hat{\beta}_1 x_k + e_k) \\ &= E_I E_R \frac{1}{n^2}\sum_{k=1}^{r}\frac{1}{p_k^2}[V_Z(\hat{\beta}_0) + V_Z(\hat{\beta}_1 x_k) + V_Z(e_k) + 2COV(\hat{\beta}_0, \hat{\beta}_1 x_k)] \\ &= \frac{1}{n^2}\sum_{k=1}^{r}\frac{1}{p_k^2}[\frac{\sigma^2\sum_{i=1}^{m}x_i^2}{m\sum_{i=1}^{m}(x_i-\hat{\mu}_{xm})^2} + \frac{\sigma^2 x_k^2}{\sum_{i=1}^{m}(x_i^2-\hat{\mu}_{xm})^2} - \frac{2x_k\hat{\mu}_{xm}\sigma^2}{\sum_{i=1}^{m}(x_i-\hat{\mu}_{xm})^2} + \sigma^2] \\ &= \frac{\sigma^2}{n^2\sum_{i=1}^{m}(x_i-\hat{\mu}_{xm})^2}\sum_{k=1}^{r}\frac{1}{p_k^2}[\frac{\sum_{i=1}^{m}x_i^2}{m} + x_i^2 - 2x_k\hat{\mu}_{xm} + \sum_{i=1}^{m}(x_i-\hat{\mu}_{xm})^2]. \end{aligned}$$

Equation 5.40 follows variance expression. Note that the regression here is weighted regression.

■

Theorem 5.21 *Random Equal Probability Sampling without Replacement under the naive model and stochastic regression imputation*

In random sampling with varying probability without replacement under the naive model and stochastic regression imputation, an unbiased estimator of the

total τ is

$$\hat{\tau}_{reg,na}^{srs,\pi ps} = [\sum_{k=1}^{m} \frac{y_k}{\pi_k} + \sum_{k=1}^{r} \frac{z_k}{\pi_k}], \tag{5.41}$$

where

$$z_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + e_k,\ \hat{\beta}_0 = \frac{\sum_{k=1}^{m} \frac{y_k}{\pi_k} - \hat{\beta}_1 \sum_{k=1}^{m} \frac{x_k}{\pi_k}}{\sum_{k=1}^{m} \frac{1}{\pi_k}} \text{ and}$$
$$\hat{\beta}_1 = \frac{\sum_{k=1}^{m} \frac{x_k y_k}{\pi_k} - (\sum_{k=1}^{m} \frac{x_k}{\pi_k} \sum_{k=1}^{m} \frac{y_k}{\pi_k}) / \sum_{k=1}^{m} \frac{1}{\pi_k}}{\sum_{k=1}^{m} \frac{x_k^2}{\pi_k} - (\sum_{k=1}^{m} \frac{x_k}{\pi_k})^2 / \sum_{k=1}^{m} \frac{1}{\pi_k}},$$

with a variance of

$$V(\hat{\tau}_{reg,na}^{srs,\pi ps}) = \sum_{k=1}^{N} \frac{1-\pi_k}{\pi_k} y_k^2 + \sum_{k=1}^{N} \sum_{j \neq k}^{N} (\frac{\pi_{kj} - \pi_k \pi_j}{\pi_k \pi_j}) y_k y_j +$$

$$\frac{S^2}{\sum_{k=1}^{m} (x_k - \hat{\mu}_{xm})^2} \sum_{k=1}^{r} \frac{1}{\pi_k^2} [\frac{\sum_{i=1}^{m} x_i^2}{m} + x_k^2 - 2 x_k \hat{\mu}_{xm} + \sum_{k=1}^{m} (x_k - \hat{\mu}_{xm})^2], \tag{5.42}$$

where $\hat{\mu}_{xm}$ is the response sample mean of the auxiliary variable X.

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that for theorem 5.20.

■

Theorem 5.22 *Stratified Equal Probability Sampling with Replacement under the naive model and stochastic regression imputation*

In stratified random sampling with varying probability with replacement under the naive model and stochastic regression imputation, an unbiased estimator of the total τ is

$$\hat{\tau}_{reg,na}^{st,pps} = \sum_{h=1}^{H} \sum_{l=1}^{L} \frac{1}{n_{hl}} [\sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{p_{hlk}} + \sum_{k=1}^{r_{hl}} \frac{z_{hlk}}{p_{hlk}}], \tag{5.43}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1} x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \frac{\sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{p_{hlk}} - \hat{\beta}_{hl1} \sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}} \text{ and}$$
$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}} \frac{x_{hlk} y_{hlk}}{p_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}} \sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{p_{hlk}}) / \sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{x_{hlk}^2}{p_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}})^2 / \sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}},$$

with a variance of

$$V(\hat{\tau}_{reg,na}^{st,pps}) = \sum_{h=1}^{H}\sum_{l=1}^{L}\{\frac{1}{n_{hl}}[\sum_{k=1}^{N_{hl}}\frac{y_{hlk}}{p_{hlk}} - \tau_{hl}^2]+$$

$$\frac{\sigma_{hl}^2}{n_{hl}^2\sum_{k=1}^{m_{hl}}(x_{hlk}-\hat{\mu}_{xhlm})^2}\sum_{k=1}^{r_{hl}}\frac{1}{p_{hlk}^2}[\frac{\sum_{i=1}^{m_{hl}}x_{hli}^2}{m_{hl}} + x_{hlk}^2 - 2x_{hlk}\hat{\mu}_{xhlm} + \sum_{i=1}^{m_{hl}}(x_{hli}-\hat{\mu}_{xhlm})^2]\}, \tag{5.44}$$

where $\hat{\mu}_{xhlm}$ is the response sample mean of the auxiliary variable X in stratum hl.

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that for theorem 5.3.

■

Theorem 5.23 *Stratified Equal Probability Sampling without Replacement Selection under the naive model and stochastic regression imputation*

In stratified random sampling with varying probability without replacement under the naive model and stochastic regression imputation, an unbiased estimator of the total τ is

$$\hat{\tau}_{reg,na}^{st,\pi ps} = \sum_{h=1}^{H}\sum_{l=1}^{L}[\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}} + \sum_{k=1}^{r_{hl}}\frac{z_{hlk}}{\pi_{hlk}}], \tag{5.45}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1}x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \frac{\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}} - \hat{\beta}_{hl1}\sum_{k=1}^{m_{hl}}\frac{x_{hlk}}{\pi_{hlk}}}{\sum_{k=1}^{m_{hl}}\frac{1}{\pi_{hlk}}} \text{ and}$$

$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}}\frac{x_{hlk}y_{hlk}}{\pi_{hlk}} - (\sum_{k=1}^{m_{hl}}\frac{x_{hlk}}{\pi_{hlk}}\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{\pi_{hlk}})/\sum_{k=1}^{m_{hl}}\frac{1}{\pi_{hlk}}}{\sum_{k=1}^{m_{hl}}\frac{x_{hlk}^2}{\pi_{hlk}} - (\sum_{k=1}^{m_{hl}}\frac{x_{hlk}}{\pi_{hlk}})^2/\sum_{k=1}^{m_{hl}}\frac{1}{\pi_{hlk}}},$$

with a variance of

$$V(\hat{\tau}_{reg,na}^{st,\pi ps}) = \sum_{h=1}^{H}\sum_{l=1}^{L}[\sum_{k=1}^{N_{hl}}\frac{1-\pi_{hlk}}{\pi_{hlk}}y_{hlk}^2 + \sum_{k=1}^{N_{hl}}\sum_{j\neq k}^{N_{hl}}(\frac{\pi_{hlkj}-\pi_{hlk}\pi_{hlj}}{\pi_{hlk}\pi_{hlj}})y_{hlk}y_{hlj}]+$$

$$\sum_{h=1}^{H}\sum_{l=1}^{L}\frac{S_{hl}^2}{\sum_{k=1}^{m_{hl}}(x_{hlk}-\hat{\mu}_{xhlm})^2}\sum_{k=1}^{r_{hl}}\frac{1}{\pi_{hlk}^2}[\frac{\sum_{i=1}^{m_{hl}}x_{hli}^2}{m_{hl}} + x_{hlk}^2 - 2x_{hlk}\hat{\mu}_{xhlm} + \sum_{k=1}^{m_{hl}}(x_{hlk}-\hat{\mu}_{xhlm})^2], \tag{5.46}$$

where $\hat{\mu}_{xhlm}$ is the response sample mean of the auxiliary variable X in stratum hl.

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that for theorem 5.3.

■

5.2.3.2 *RHG* Model in Stochastic Regression Imputation

I extend and prove stochastic regression imputation with the *RHG* model in theorems 5.24-5.27. To obtain the estimated variance for equal probability sampling design in theorems 5.24 and 5.25 replace σ^2_{hl} for sampling with replacement or S^2_{hl} for sampling without replacement with $s^2_{hlm} = \frac{1}{m_{hl}-1}\sum_{k=1}^{m_{hl}}(y_{hlk} - \hat{\mu}_{hlm})^2$, where $\hat{\mu}_{hlm} = \frac{1}{m_{hl}}\sum_{k=1}^{m_{hl}} y_{hlk}$ and m_{hl} is the response size in the post-stratum hl. For remaining theorems, estimates of variance are made as in theorem 2.7 or 2.8.

Theorem 5.24 *Simple random sampling with replacement under the RHG model and stochastic regression imputation*

In simple random sampling with replacement under the *RHG* model and stochastic regression imputation, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}^{srs}_{reg,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nn_{hl}}[\sum_{k=1}^{m_{hl}} y_{hlk} + \sum_{k=1}^{r_{hl}} z_{hlk}], \tag{5.47}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1}x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \hat{\mu}_{yhlm} - \hat{\beta}_{hl1}\hat{\mu}_{xhlm} \text{ and}$$
$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})(y_{hlk} - \hat{\mu}_{yhlm})}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2},$$

with a conditional variance of

$$V(\hat{\mu}^{srs}_{reg,RHG}|\mathbf{n},\mathbf{m}) =$$

$$\sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2[\frac{\sigma_{hl}^2}{n_{hl}}+(\frac{r_{hl}\sigma_{hl}}{n_{hl}})^2\{\frac{(\hat{\mu}_{xhlm}-\hat{\mu}_{xhlr})^2}{\sum_{k=1}^{m_{hl}}(x_{hlk}-\hat{\mu}_{xhlm})^2}+\frac{1}{m_{hl}}+\frac{1}{r_{hl}}\}], \tag{5.48}$$

where $\hat{\mu}_{xhlm}$ and $\hat{\mu}_{xhlr}$ are the response and nonresponse sample mean of the auxiliary variable X in stratum hl respectively and $\hat{\mu}_{yhlm}$ is the response sample mean of the study variable Y in stratum hl.

Proof: The sample is post-stratified into $H_S \times L_S$ post-strata. By theorem 5.9,

$$\begin{aligned} E(\hat{\mu}_{reg,RHG}^{srs}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}E(\hat{\mu}_{hl,reg,na}^{srs}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{N}\mu_{hl} \\ &= \mu. \end{aligned}$$

The variance of $\hat{\mu}$ can be expressed as

$$V(\hat{\mu}|\mathbf{n},\mathbf{m}) = V_I E_R E_Z(\hat{\mu}|Y,X,Z,I,R,\mathbf{n},\mathbf{m}) + E_I V_R E_Z(\hat{\mu}|Y,X,Z,I,R,\mathbf{n},\mathbf{m}) + E_I E_R V_Z(\hat{\mu}|Y,X,Z,I,R,\mathbf{n},\mathbf{m}).$$

To prove equation 5.48 note that

$$\begin{aligned} V_I E_R E_Z(\hat{\mu}_{reg,RHG}^{srs}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V_1(\hat{\mu}_{yhl}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{\sigma_{hl}^2}{n_{hl}}, \end{aligned}$$

$$\begin{aligned} E_I V_R E_Z(\hat{\mu}_{reg,RHG}^{srs}|\mathbf{n},\mathbf{m}) &= \frac{1}{n}E_I V_R[\sum_{k=1}^{N} I_k\frac{y_k}{p_k}] \\ &= 0, \end{aligned}$$

and

$$\begin{aligned} E_I E_R V_Z(\hat{\mu}_{reg,RHG}^{srs}|\mathbf{n},\mathbf{m}) &= E_I V_J \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{n_{hl}N}(\sum_{k=1}^{r_{hl}}(\hat{\beta}_0+\hat{\beta}_1 x_k+e_k)) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2(\frac{r_{hl}\sigma_{hl}}{n_{hl}})^2[\frac{(\hat{\mu}_{xhlm}-\hat{\mu}_{xhlr})^2}{\sum_{h=1}^{m_{hl}}(x_{hlk}-\hat{\mu}_{xhlm})^2}+\frac{1}{m_{hl}}+\frac{1}{r_{hl}}]. \end{aligned}$$

Equation 5.48 follows by the variance expression.

■

Theorem 5.25 *Simple random sampling without replacement under the RHG model and stochastic regression imputation*

In Simple random sampling without replacement under the *RHG* model and stochastic regression imputation, a conditional unbiased estimator of the mean μ is

$$\hat{\mu}^{srs}_{reg,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{N_{hl}}{Nn_{hl}}[\sum_{k=1}^{m_{hl}} y_{hlk} + \sum_{k=1}^{r_{hl}} z_{hlk}], \tag{5.49}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1}x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \hat{\mu}_{yhlm} - \hat{\beta}_{hl1}\hat{\mu}_{xhlm} \text{ and}$$
$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})(y_{hlk} - \hat{\mu}_{yhlm})}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2},$$

with a conditional variance of

$$V(\hat{\mu}^{srs}_{reg,RHG}|\mathbf{n},\mathbf{m})$$
$$= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2[\frac{N_{hl}-n_{hl}}{n_{hl}N_{hl}}S^2_{hl} + (\frac{r_{hl}S_{hl}}{n_{hl}})^2\{\frac{(\hat{\mu}_{xhlm} - \hat{\mu}_{xhlr})^2}{\sum_{k=1}^{m_{hl}}(x_{hlk} - \hat{\mu}_{xhlm})^2} + \frac{1}{m_{hl}} + \frac{1}{r_{hl}}\}], \tag{5.50}$$

where $\hat{\mu}_{xhlm}$ and $\hat{\mu}_{xhlr}$ are the response and nonresponse sample mean of the auxiliary variable X in stratum hl respectively and $\hat{\mu}_{yhlm}$ is the response sample mean of the study variable Y in stratum hl.

Proof: Proof can be followed as theorem 5.24 except

$$\begin{aligned} V_I E_R E_Z(\bar{y}^{srs}_{reg,RHG}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2 V_1(\bar{y}_{hl}) \\ &= \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}(\frac{N_{hl}}{N})^2\frac{N_{hl}-n_{hl}}{n_{hl}N_{hl}}S^2_{hl} \end{aligned}$$

■

Theorem 5.26 *Random Equal Probability Sampling with Replacement under the RHG model and stochastic regression imputation*

In random sampling with varying probability with replacement under the *RHG* model and stochastic regression imputation, a conditional unbiased estimator of τ is

$$\hat{\tau}^{srs,pps}_{reg,RHG} = \sum_{h=1}^{H_S}\sum_{l=1}^{L_S}\frac{1}{n_{hl}}[\sum_{k=1}^{m_{hl}}\frac{y_{hlk}}{p_{hlk}} + \sum_{k=1}^{r_{hl}}\frac{z_{hlk}}{p_{hlk}}], \tag{5.51}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1} x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \frac{\sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{p_{hlk}} - \hat{\beta}_{hl1} \sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}} \text{ and}$$

$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}} \frac{x_{hlk} y_{hlk}}{p_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}} \sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{p_{hlk}}) / \sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{x_{hlk}^2}{p_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{p_{hlk}})^2 / \sum_{k=1}^{m_{hl}} \frac{1}{p_{hlk}}},$$

with a conditional variance of

$$V(\hat{\tau}_{reg,RHG}^{srs,pps}|\mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S} \sum_{l=1}^{L_S} \{ \frac{1}{n_{hl}} [\sum_{k=1}^{N_{hl}} \frac{y_{hlk}}{p_{hlk}} - \tau_{hl}^2] +$$

$$\frac{\sigma_{hl}^2}{n_{hl}^2 \sum_{k=1}^{m_{hl}} (x_{hlk} - \bar{x}_{hlm})^2} \sum_{k=1}^{r_{hl}} \frac{1}{p_{hlk}^2} [\frac{\sum_{i=1}^{m_{hl}} x_{hli}^2}{m_{hl}} + x_{hlk}^2 - 2 x_{hlk} \bar{x}_{hlm} + \sum_{i=1}^{m_{hl}} (x_{hli} - \bar{x}_{hlm})^2] \}, \tag{5.52}$$

where $\hat{\mu}_{xhlm}$ is the response sample mean of the auxiliary variable X in stratum hl.

Proof: The sample is post-stratified into $H_S \times L_S$ post-strata. By theorem 5.20,

$$\begin{aligned} E(\hat{\tau}_{reg,RHG}^{srs,pps}|\mathbf{n},\mathbf{m}) &= \sum_{h=1}^{H_S} \sum_{l=1}^{L_S} E(\hat{\tau}_{hl,reg,na}^{srs,pps}) \\ &= \sum_{h=1}^{H_S} \sum_{l=1}^{L_S} \tau_{hl} \\ &= \tau. \end{aligned}$$

By the results in theorem 5.20, a conditional variance of the total estimator is in equation 5.52.

■

Theorem 5.27 *Random Equal Probability Sampling without Replacement under RHG model and stochastic regression imputation*

In random sampling with varying probability without replacement under *RHG* model and stochastic regression imputation, a conditional unbiased estimator of τ is

$$\hat{\tau}_{reg,RHG}^{srs,\pi ps} = \sum_{h=1}^{H_S} \sum_{l=1}^{L_S} [\sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{\pi_{hlk}} + \sum_{k=1}^{r_{hl}} \frac{z_{hlk}}{\pi_{hlk}}], \tag{5.53}$$

where

$$z_{hlk} = \hat{\beta}_{hl0} + \hat{\beta}_{hl1} x_{hlk} + e_{hlk},\ \hat{\beta}_{hl0} = \frac{\sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{\pi_{hlk}} - \hat{\beta}_{hl1} \sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{\pi_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{1}{\pi_{hlk}}} \text{ and}$$

$$\hat{\beta}_{hl1} = \frac{\sum_{k=1}^{m_{hl}} \frac{x_{hlk} y_{hlk}}{\pi_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{\pi_{hlk}} \sum_{k=1}^{m_{hl}} \frac{y_{hlk}}{\pi_{hlk}}) / \sum_{k=1}^{m_{hl}} \frac{1}{\pi_{hlk}}}{\sum_{k=1}^{m_{hl}} \frac{x_{hlk}^2}{\pi_{hlk}} - (\sum_{k=1}^{m_{hl}} \frac{x_{hlk}}{\pi_{hlk}})^2 / \sum_{k=1}^{m_{hl}} \frac{1}{\pi_{hlk}}},$$

with a conditional variance of

$$V(\hat{\tau}_{reg,RHG}^{srs,\pi ps} | \mathbf{n},\mathbf{m}) = \sum_{h=1}^{H_S} \sum_{l=1}^{L_S} [\sum_{k=1}^{N_{hl}} \frac{1-\pi_{hlk}}{\pi_{hlk}} y_{hlk}^2 + \sum_{k=1}^{N_{hl}} \sum_{j \neq k}^{N_{hl}} (\frac{\pi_{hlkj} - \pi_{hlk}\pi_{hlj}}{\pi_{hlk}\pi_{hlj}}) y_{hlk} y_{hlj}\}] +$$

$$\sum_{h=1}^{H_S} \sum_{l=1}^{L_S} \frac{S_{hl}^2}{\sum_{k=1}^{m_{hl}} (x_{hlk} - \hat{\mu}_{xhlm})^2} \sum_{k=1}^{r_{hl}} \frac{1}{\pi_{hlk}^2} [\frac{\sum_{i=1}^{m_{hl}} x_{hli}^2}{m_{hl}} + x_{hlk}^2 - 2 x_{hlk} \hat{\mu}_{xhlm} + \sum_{i=1}^{m_{hl}} (x_{hli} - \hat{\mu}_{xhlm})^2], \tag{5.54}$$

where $\hat{\mu}_{xhlm}$ is the response sample mean of the auxiliary variable X in stratum hl.

Proof: Proof for unbiasedness of the total estimator and for its variance is similar to that for theorem 5.26.

■

5.3 Multiple Imputation

A very thorough introduction to multiple imputation is presented by *Barnard et al* (1998). The outline of this is:

> Although single imputation satisfies critical data-processing objectives and can incorporate knowledge from the data procedure, it fails to satisfy statistical objectives concerning the validity of the resulting inferences based on the completed data. Specifically, for validity, the resulting estimates based on the data completed by imputation should be approximately unbiased for their population estimates, confidence intervals should attain at least their nominal coverages, and tests of null hypotheses should not reject true null hypotheses more frequently than their nominal levels. Because a single imputed value cannot reflect any of the uncertainty

about the true underlying values, analyses that treat imputed value just like observed values underestimate uncertainty. Thus, imputing a single value for each missing datum and then analysing the completed data will result in standard error estimates that are too small, confidence intervals that fail to attain their nominal coverages, and P values that are too significant; this is true even if the modelling for imputation is carried out carefully.

Multiple imputation, first proposed in *Rubin* (1978), is an approach that retains the advantages of single imputation while allowing the data analyst to obtain valid assessments of uncertainty. The basic idea is to impute two or more times for the missing data using independent draws of the missing values from a distribution that is appropriate under the postulations about the data and the mechanism creating missing data. This results in two or more completed data sets, each of which is analysed using the same standard complete-data method. The analyses are then combined in a simple way that reflects the extra uncertainty due to having imputed. Multiple imputations may also be created under several different models to display sensitivity to the choice of missing-data model.

Theoretical motivation for multiple imputation is described in section 5.3.1. Section 5.3.2 deals with analysis of a multiply imputed data set. Ignorable nonresponse imputation and nonignorable nonresponse imputation techniques are described in sections 5.3.3 and section 5.3.4 respectively.

5.3.1 Theoretical Motivation for Multiple Imputation

As discussed by *Barnard et al* (1998), the theoretical motivation for multiple imputation is Bayesian, although the procedure has excellent properties from a frequentist perspective. More information on the properties of multiple imputation are in *Rubin* (1987, 1996), *Herzog & Rubin* (1983, 1986), *Li et al* (1991), *Rubin & Schenker* (1987)

and *Meng & Rubin* (1992), *Heitjan & Rubin* (1990), *Nordholt* (1998), *Schenker & Taylor* (1996), *Schenker & Welsh* (1988), *Sedransk et al* (1991), *Sedransk & Jinn* (1992), *Shao & Sitter* (1996), *Zeger & Karim*(1991) and *Schafer* (1997).

Formally, let Q be the population quantity of interest, and suppose the data can be partitioned into observed values, $\mathbf{Y}_{obs}$, and missing values, $\mathbf{Y}_{mis}$. If $\mathbf{Y}_{mis}$ had been observed, then inferences for Q would have been based on the complete-data posterior density $p(Q|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$. Because $\mathbf{Y}_{mis}$ is not observed, inferences are based on the actual posterior density $p(Q|\mathbf{Y}_{obs})$, which can be expressed as

$$p(Q|\mathbf{Y}_{obs}) = \int p(Q|\mathbf{Y}_{obs}, \mathbf{Y}_{mis})p(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})d\mathbf{Y}_{mis} \tag{5.55}$$

Equation 5.55 shows that the actual posterior density of Q can be obtained by averaging the complete-data posterior density over the posterior predictive distribution of $\mathbf{Y}_{mis}$. In practice, multiple imputations are repeated independent draws from $p(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$. Thus, multiple imputation allows the data analyst to approximate (5.55) by separately analysing each data set completed by imputation and then combining the results of the separate analyses.

5.3.2 Analysing a Multiply Imputed Data Set

The exact computation of the posterior distribution (5.55) by simulation would require that an infinite number of values of $\mathbf{Y}_{mis}$ be drawn from $p(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$. If the data were complete, inferences for Q would be based on a point estimate $\hat{Q}$, an associated variance estimate $\hat{U}$, and a normal reference distribution. When data are missing and there are M sets of imputations for the missing data, M sets of complete-data statistics are $\hat{Q}_l$ and $\hat{U}_l$ for $l=1,...,M$.

Rubin & Schenker (1986) suggest the following procedure for drawing inferences about Q from the multiply imputed data. The point estimate of Q is the average of the M complete-data estimates,

$$\bar{Q} = \sum_{l=1}^{M} \frac{\hat{Q}_l}{M},$$

and the associated variance estimate is

$$T = \bar{U} + (1 + M^{-1})B,$$

where $\bar{U} = \sum_{l=1}^{M} \frac{\hat{U}_l}{M}$ is the average within-imputation variance, and $B = \frac{1}{M-1} \sum_{l=1}^{M} (\hat{Q}_l - \bar{Q})^2$ is the between-imputation variance. The approximate reference distribution for interval estimates and hypothesis testings is a t distribution with degrees of freedom

$$\nu = (M - 1)(1 + r^{-1})^2,$$

where $r = (1 + M^{-1})\frac{B}{\bar{U}}$ is the estimated ratio of the between-imputation component of variance to the average within-imputation component of variance.

Ideally, multiple imputations are M independent random draws from the posterior predictive distribution of $\mathbf{Y}_{mis}$ under appropriate Bayesian modelling assumptions. In practice, approximations to the posterior distribution are often used and work well. Such imputations are called repeated imputation in *Rubin* (1987).

Several important issues arise in the creation of imputation models. These include explicit vs. implicit models, and ignorable vs. nonignorable models.

Imputation procedures can be based on explicit models or implicit models, or even combinations (*Rubin*, 1987). An example of a procedure based on explicit model is regression imputation. This method uses respondent data to regress the variable for which imputations are required on an auxiliary variable, X. The regression equation is then used to predict the values for the missing data. The imputed value may either be the predicted value, or the predicted value plus some residual. A common type of procedure based on implicit models is hot-deck imputation, which replaces the missing values for an incomplete case, where the matching is carried

out with respect to variables that are observed for both the incomplete cases and complete cases.

The model underlying an imputation procedure, whether explicit or implicit, can be based on the assumption that the reasons for missing data are either ignorable or nonignorable (*Rubin*, 1976). The distinction between an ignorable and a nonignorable model are discussed in section 2.6.3. An important issue with nonignorable models is that because the missing values cannot be observed, there is no direct evidence in the data to address the assumption of nonignorability. It can be important, therefore, to consider several alternative models and to explore a sensitivity analysis of resulting inferences to the choice of model. *Rubin* (1976) point out imputation methods:

> In current practice, almost all imputation models are assumed to be ignorable; limited experience suggests that in major surveys with limited amounts of missing data and careful design, ignorable models are satisfactory for most analyses.

5.3.3 Ignorable Nonresponse Techniques

Rubin & Schenker (1986) presents multiple imputation methods with ignorable nonresponse. These methods assume that the response mechanism generating the missing data is ignorable. Random, Bayesian bootstrap and approximate Bayesian bootstrap, fully normal and adjusted fully normal imputation methods are proposed. *Rubin* (1987) also proposes the hot-deck imputation method such as sequential imputation and regression method for ignorable nonresponse techniques.

There are six ignorable nonresponse techniques studied in this thesis: random, sequential, stochastic regression, approximate Bayesian bootstrap, fully normal and adjusted fully normal imputation methods.

In this section multiple imputation theorems are not stated but the way to draw the imputed data from ignorable nonresponse technique procedures are pre-

sented. Random, sequential and stochastic regression methods are used those procedures from single imputation methods in section 5.2. Procedures with approximate Bayesian bootstrap, fully normal and adjusted fully normal imputation methods are presented in section 5.3.3.1-5.3.3.3 respectively. Before these procedures are presented, an example is given on how estimated mean and its variance for these multiple imputation methods are calculated by using formulae in section 5.3.2.

For example, if multiple random imputation is used for compensating missing data in *SRSWOR* (*Govindarajulu*, 1999), the multiple imputation estimate of μ is the average of the M complete-data estimates of μ,

$$\begin{aligned}\hat{\mu}_{ran,na}^{srs,M} &= \frac{1}{M}\sum_{l=1}^{M}\hat{\mu}_{ran,na,l}^{srs} \\ &= \frac{1}{M}\sum_{l=1}^{M}\frac{(k+1)m\hat{\mu}_m + t\hat{\mu}_{t,l}}{n}\end{aligned}$$

and its estimated variance is given by

$$V(\hat{\mu}_{ran,na}^{srs,M}) = \frac{1}{M}\sum_{l=1}^{M}[(\frac{1}{m}-\frac{1}{N})s_{m,l}^2 + \frac{t}{n}(1-\frac{t}{m})\frac{s_{m,l}^2}{n}] + \frac{M+1}{M(M-1)}\sum_{l=1}^{M}(\hat{\mu}_{ran,na,l}^{srs} - \hat{\mu}_{ran,na}^{srs,M})^2$$

5.3.3.1 Approximate Bayesian Bootstrap Imputation

Rubin & Schenker (1986) give the brief idea with Bayesian Bootstrap (BB) imputation as follow:

> Suppose each element of the population takes one of the values $d_1, ..., d_K$ with probabilities $\theta_1, ..., \theta_K$, respectively. It the improper Dirichlet prior with density proportional to $\Pi_{k=1}^{K}\theta_k^{-1}$ is placed on the vector $\theta = (\theta_1, ..., \theta_K)$, then the posterior distribution of θ is the Dirichlet distribution with density proportional to $\Pi_{k=1}^{K}\theta_k^{q_k-1}$ and $K-$dimensional mean vector $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_K)$ having components given by $\hat{\theta}_k = q_k/n_1$, where q_1 is the number of times d_k appears in Y_{obs}. Components of the responding to values of d_k not appearing in Y_{obs} will be with probability

one. The *BB* method first observed a value θ^* of θ from this posterior distribution. Then the components of Y_{mis} are independently drawn from among $d_1, ..., d_K$ using the probabilities θ^*.

Rubin & Schenker (1986) suggest a simple approximation to the Bayesian Bootstrap (ABB) method that is more direct from the computational point of view. The ABB procedures are summarised as follow:

Step 1. Draw the data with replacement from Y_{obs} m times, $\mathbf{D} = (D_1, ..., D_m)$.

Step 2. Draw the $r = n - m$ components of Y_{mis} with replacement from $\mathbf{D}$ in step 1.

Step 3. Repeat steps (1)-(2) M times to form the M multiple data sets.

The only difference between ABB and BB methods are that instead of drawing θ from the Dirichlet posterior distribution as in the BB method, the ABB method draws θ from a scaled multinomial distribution. The distribution used for θ in the BB and ABB methods have the same mean vectors and the same correlations; however, the variances for the ABB method are $\frac{m+1}{m}$ times the variances for the BB method. More details see *Rubin* (1981).

5.3.3.2 Fully Normal Imputation

Assume that the data is a random sample from a $\mathrm{N}(\mu, \sigma^2)$ distribution. If the prior distribution of (μ, σ^2) has density proportional to σ^{-2}, then the posterior distribution of σ^2 is $(m-1)\frac{s_m^2}{\chi^2_{m-1}}$, and the conditional posterior distribution of μ given σ^2 is $\mathrm{N}(\hat{\mu}_m, \frac{\sigma^2}{m})$ (*Box & Tiao*, 1973).

Rubin & Schenker (1986) summarise the fully normal (FN) procedures as follow:

Step 1. Draw a value (μ^*, σ^{*2}) of (μ, σ^2) from the posterior distribution of (μ, σ^2): σ^{*2} is drawn from $(m-1)\frac{s_m^2}{\chi^2_{m-1}}$.

Step 2. Draw μ^* from $\mathrm{N}(\hat{\mu}_m, \frac{\sigma^{*2}}{m})$.

Step 3. The $r = n-m$ components of Y_{mis} are then drawn as a random sample from $\mathrm{N}(\mu^*, \sigma^{*2})$ population.

Step 4. Repeat steps (1)-(3) M times to form the M multiple data sets.

5.3.3.3 Adjusted Fully Normal Imputation or Imputation adjusted for Uncertainty in the Mean and Variance

Rubin & Schenker (1986) describe the normality assumption for adjusted fully normal imputation:

> If the assumption of normality is not valid, the observed data Y_{obs} is desirable assumed to influence the shape of the distribution of imputed values for Y_{mis}. The adjusted fully normal imputation (AFN) method achieves this idea because the AFN method draws values from Y_{obs} in place of simulated normal drawn in the FN method, the shape of the distribution of imputed values is influenced by Y_{obs}. For example, if the values in Y_{obs} are left-skewed, then the AFN method will lead to left-skewed imputation for Y_{mis}. On the other hand, the FN method will lead to symmetric imputation for Y_{mis} irrespective of the distribution of values in Y_{obs}

The AFN algorithms are described as follow:

Ste 1. μ^* and σ^{*2} are drawn as in the FN method.

Step 2. The $r = n - m$ components of Y_{mis} are then drawn with replacement from Y_{obs}, $\mathbf{D} = (D_1, D_2, ..., D_r)$.

Step 3. Compute $Z_k = (D_k - \hat{\mu}_m)[\frac{(m-1)s_m^2}{m}]^{0.5}$ which has expected value zero and variance 1 under repeated draws from Y_{obs}.

Step 4. Compute the missing value $Y_k = \mu^* + \sigma^* Z_k, k = 1, ..., r$.

Step 5. Repeat steps (1)-(4) M times to form the M multiple data sets.

5.3.4 Nonignorable Nonresponse Techniques

In practice almost all surveys suffer from nonresponse. When response is unrelated to the values of the missing variable Y, the nonresponse is called ignorable (*Little*, 1982). Multiple imputation methods in section 5.3.3 above were described for ignorable nonresponse techniques. However, in many cases the nonresponse could be related to the value of variable Y. This may be the case, for example, in surveys of income, of alcohol consumption, and of injuries where litigation is possible (*Glynn et al*, 1993). This type of nonresponse is called nonignorable.

Tanner (1996) summarises the ideas for using nonignorable nonresponse techniques from *Rubin* (1987). These leads to two approaches to handling nonignorable nonresponse with implicit models: *mixture models* and *selection models*. Each of these two approaches can be use in the absence or presence of follow-up data. Only mixture models are used in this thesis because selection models have been used with response probability functions and this leads to more complications in simulation.

Wang et al (1992) study the performance of confidence intervals for simple linear regression coefficients based on multiple imputation for explicit model when missing values cannot be regarded as *MAR*.

In this thesis, I review the theory of multiple imputation on a mixture model without follow-up data in section 5.4.4.1 and with follow-up data in section 5.4.4.2. Modified *Wang's* regression algorithm to compute the mean estimate is presented in section 5.4.4.3.

To estimate the mean and its variance for nonignorable nonresponse, the same formulae in section 5.3.2 is used.

5.3.4.1 Mixture Model Without Follow-up Data

Rubin (1987) factors the posterior distribution of the missing values and the parameters of the model into three components. The first component is the predictive distribution of the missing data given the parameters. The second component is the conditional distribution of the nonresponders' parameters given the responders' parameters. The third component is the posterior of the responders' parameters given the observed data having specified each of these distributions.

As discussed with consideration between a response Y and a covariate X by *Tanner*(1996), the normal mixture model supposes that for the responders, Y_k follows the normal distribution with mean $\alpha_m + \beta_m X_k$ and variance σ_m^2. For the nonresponders, Y_k follows the normal distribution with mean $\alpha_r + \beta_r X_k$ and variance σ_r^2. For simplicity it is assumed $\sigma_m = \sigma_r = \sigma$. Hence, under a noninformative prior, the posterior distribution $p(\alpha_m, \beta_m, \sigma|Y)$ factors as the product of an inverse chi-square distribution and a conditional normal distribution. Note that in large samples this factorisation should be approximately correct even if the Y_k is not normally distributed. The predictive distribution of the missing data, $p(Z_k|\alpha_m, \beta_m, \alpha_r, \beta_r, \sigma, Y)$, is $N(\alpha_r + \beta_r X_k, \sigma^2)$. *Rubin* (1987) assumes that X_k is observed for all units in the sample.

The conditional distribution of β_r and α_r given α_m, β_m and σ is given by the product of two independent distributions. The conditional marginal of β_r given α_m, β_m and σ is a normal distribution with mean β_m and variance $C_\beta^2 \beta_m^2$, where C_β^2 is the prior coeffiecient of variation in the slope of regression that specifies the similarity of slopes for responders and nonresponders. The conditional marginal of α_r is specified through the average Y value at $\bar{X}_m$ for the nonresponders in the population, $\eta_r = \alpha_r + \beta_r \bar{X}_m$. The conditional marginal of η_r is normal with mean $\eta_m = \alpha_m + \beta_m \bar{X}_m$ and variance $C_\eta^2 \eta_m^2$, where C_η^2 is the prior coeffiecient of variation in intercept of regression that specifies the similarity of the expected value of Y for the responders and nonresponders with covariate mean equal to the covariate mean of the responders. Note that when $C_\eta = C_\beta = 0$ these specifications imply an

ignorable nonresponse mechanism.

In this way, the joint posterior of the missing data $Z, \eta_r, \beta_r, \eta_m, \beta_m$ and σ factors as:

$$p(Z|\eta_r, \beta_r, \sigma, Y)p(\eta_r, \beta_r|\eta_m, \beta_m, \sigma, Y)p(\eta_m, \beta_m, \sigma|Y), \tag{5.56}$$

where the components are defined above.

To implement this normal mixture model, *Rubin* (1987) performs the following three steps to impute the missing data:

Step 1. Draw η_m^*, β_m^* and σ_*^2.

1.1) Draw σ_*^2 from the inverse χ^2 distribution: $\frac{(m-2)s_m^2}{\chi^2_{(m-2)}}$, where s_m^2 is the residual mean square for the responders' regression.

1.2) Draw β_m^* from the conditional normal: $N(\hat{\beta}_m, \sigma_*^2(\sum_{k=1}^m x_k^2)^{-1})$, where $\hat{\beta}_m$ is the least-squares estimate of β.

1.3) Draw η_m^* from $N(\hat{\mu}_m, \frac{\sigma_*^2}{m})$.

Step 2. Draw β_r^* and η_r^*.

2.1) Draw β_r^* from $N(\beta_m^*, C_\beta^2 \beta_m^{*2})$.

2.2) Draw η_r^* from $N(\eta_m^*, C_\eta^2 \eta_m^{*2})$.

Step 3. Draw $r = n - m$ imputed missing values Z from $N(\eta_r^* + (X_k - \bar{X}_m)^2 \beta_r^*, \sigma_*^2)$.

Step 4. Repeat steps (1)-(3) M times to form the M multiple data sets.

5.3.4.2 Mixture Model With Follow-up Data

The only way to reduce sensitivity of inference for nonignorable nonresponse is to reduce nonresponse or accumulate information about how nonrespondents differ from respondents on the outcome variables under investigation. The most direct method for accumulating information on nonrespondents is to follow up at least

some of them to obtain the desired information. Even if only a few nonrespondents are followed up, these can be exceedingly helpful in reducing sensitivity of inference.

Glynn et al (1986) use multiple imputation to draw inferences from survey data of retired men with follow-ups using an extension of the mixture model that includes covariates.

Glynn et al (1993) approach inference for means or linear regression parameters in mixture modelling when the outcomes is subject to nonignorable nonresponse. Mixture models assume separate parameters for respondents and nonrespondents; implementation by multiple imputation consists of repeatedly filling in missing values for nonrespondents, estimating parameters using the filled-in data, and then adjusting for variability between imputations. The performance of this scheme is evaluated by using simulated data with a 25% sample of nonrespondents followed up.

Tanner (1996) summaries the the theory of mixture model with follow-up data from *Rubin* (1987) as:

> In the presence of follow-ups data, it is assumed that
>
> $$Y = \alpha_r + \beta_r X_f + \sigma\varepsilon, \tag{5.57}$$
>
> where $\varepsilon \sim N(0,1)$ and X_f is the X variable for the follow-ups. It can be factorised as
>
> $$p(Z, \beta_r, \alpha_r, \sigma|Y) = p(Z|\beta_r, \alpha_r, \sigma, Y)p(\sigma, \beta_r, \alpha_r|Y), \tag{5.58}$$
>
> where Z is a missing data.
>
> Under the noninformative prior, $p(\sigma, \beta_r, \alpha_r|Y)$ factors as $\frac{s_f^2(n_f-2)}{\chi^2_{(n_f-2)}}$ times the conditional normal $N(\hat{\beta}_r, \sigma^2(\sum_{k=1}^{n_f} x_k^2)^{-1})$, where n_f is the number of follow-ups, $\hat{\beta}_r$ is the least-squares estimate using the follow-up data and s_f^2 is the corresponding residual mean square. Note that in large samples this factorisation should be approximately correct even if the $Y_k's$ are not normally distributed. This approach

does assume that the follow-up data are a random sample from the nonresponders. The predictive distribution is $N(\alpha_r + \beta_r X_{n_f}, \sigma^2)$, where X_{n_f} is the variable X for the nonfollow-ups.

The following algorithms under the method of composition to (5.58) are summarised to yields a missing data as follow:

Step 1. Draw α_r^*, β_r^* and σ_*^2 from the inverse chi-squared $\frac{s_f^2(n_f-2)}{\chi^2_{(n_f-2)}}$ times conditional normal $N(\hat{\beta}_r, \sigma^2(\sum_{k=1}^{n_f} x_k^2)^{-1})$.

Step 2. Draw the imputed data for the nonfollow-ups from $N(\alpha_r^* + \beta_r^* X_{n_f}, \sigma_*^2)$.

Step 3. Repeat steps (1)-(2) M times to form the M multiple data sets.

The predictive distribution can alternatively be approximated using a hot-deck type approach on the follow-up data. This approach would eliminate the need to specify a normal distribution on the nonfollow-up responses (*Rubin*, 1987).

5.3.4.3 Modified Wang's Regression Method

Assume throughout the missing data cannot be regarded as *MAR*. *Wang et al* (1992) study the actual coverage probabilities of confidence intervals of the slope and intercept in a linear regression. That is, they assume

$$Y_k = \beta_0 + \beta_1 x_k + \epsilon_k, k = 1, ..., n,$$

where the ϵ_k are independent with $\epsilon_k \sim N(0, \sigma^2)$. The value of the auxiliary variable X is assumed to be known for each unit and the probability of a response on Y is

$$g(R|y, x) = \alpha_1\{1 - \alpha_1 exp(-\alpha_3 y)\}.$$

To simplify the analysis, the special case of $g(R|y, x)$ is used as

$$g(R|Y, X) = 1 - exp(-\alpha y).$$

Wang et al (1992) use several multiple imputations, e.g., overall mean imputation, overall random imputation deterministic regression imputation and stochastic regression imputation, incorporating knowledge about the nonresponse process. The imputations are obtained from the unconditional distribution of the values of Y corresponding to the nonresponding units given the sample data. To simplify the case, σ^2 is assumed known and take a noninformative prior distribution of β. Assuming that $y_1, ..., y_m$ are responded and $y_{m+1}, ..., y_n$ are nonresponded. Thus, the imputations $y_{m+1}, ..., y_n$ are taken from the density

$$f(y_{m+1}, ..., y_n | y_1, ..., y_m; x_1, ..., x_n; R_1, ..., R_n; \sigma^2) =$$
$$\int \cdots \int f_1(y_{m+1}, ..., y_n | \beta, \sigma^2, x_1, ..., x_n; R_1, ..., R_n) \times$$

$$f_2(\beta | y_1, ..., y_m; x_1, ..., x_n; R_1, ..., R_n; \sigma^2) d\beta, \tag{5.59}$$

where $R_1, ..., R_m = 1$ and $R_{m+1}, ..., R_n = 0$ are the indicator functions for the respondents and the nonrespondents, respectively.

As seen from (5.59), one of the components of this unconditional distribution is posterior distribution, f_2 of $\beta = (\beta_0, \beta_1)^T$. The distribution f_2 is proportional to the likelihood of β given the data. *Rubin* (1987) has shown that the following procedure provides confidence interval with (approximately) correct coverage probabilities. The distribution f_2 in (5.59) is proportional to

$$K\{exp[-\alpha\{\textstyle\sum_{k=m+1}^{n}(\beta_0 + \beta_1 x_k)\} - (2\sigma^2)^{-1}\{\sum_{k=1}^{m}(y_k - \beta_0 - \beta_1 x_k)^2\}]\},$$

where $K = \Pi_{k=m+1}^{n} \phi\{(\beta_0 + \beta_1 x_k)\sigma^{-1} - \alpha\sigma\}\phi((\beta_0 + \beta_1 x_k)\sigma^{-1})$ and ϕ denotes the distribution function of a standard normal variable. In this case, $(\beta_0 + \beta_1 x_k)\sigma^{-1}$ is sufficiently large so that one can set $K \doteq 1$ and f_2 can be reasonably approximated by the bivariate normal distribution. Thus,

$$f_2(\beta | y_1, ..., y_m; x_1, ..., x_n; R_1, ..., R_n; \sigma^2) \sim N(\eta, \Sigma), \tag{5.60}$$

where

$$X_m' = \begin{pmatrix} 1, & \ldots, & 1 \\ x_1, & \ldots, & x_m \end{pmatrix},$$

$$X_r' = \begin{pmatrix} 1, & \ldots, & 1 \\ x_{m+1}, & \ldots, & x_n \end{pmatrix},$$

$Y_m' = (y_1, ..., y_m)$, $\hat{\beta}_m = (X_m^T X_m)^{-1} X_m^T Y_m$, $A = (X_m^T X_m)/\sigma^2$, $\Sigma = A^{-1}$, $C = 2\alpha \mathbf{1}^T X_r^T$ with $\mathbf{1}^T = (1, ..., 1)$ a column vector of size $r = n - m$ and $\eta = (C - 2\hat{\beta}_m A)A^{-1}/2$.

In order to obtain an imputed data from (5.60) the following procedures are summaried as follow:

Step 1. Select $(\beta_0, \beta_1) = \beta_0^*, \beta_1^*$ from f_2 in (5.60).

Step 2. Given β_0^*, β_1^*, choose $\hat{y}_j$ from $(y_j | x_j, \beta_0^*, \beta_1^*, \sigma^2) \sim N(\beta_0^* + \beta_1^* x_j, \sigma^2)$.

Step 3. Select a random number u_j from the uniform (0,1) distribution.

Step 4. If $u_j \leq exp\{-\alpha \hat{y}_j\}$, then $y_j = \hat{y}_j$. Otherwise, repeat steps 2-4 until an acceptable value of y_j is found.

Step 5. Repeat steps 2-4 independently for $j = m + 1, ..., n$.

Step 6. Using $\hat{y}_{m+1}, ..., \hat{y}_n$ and the observed data values $y_1, ..., y_m$, then calculate

$$\hat{\beta}_{1c} = \frac{\sum_{k=1}^n y_k^+ (x_k - \bar{x}_n)}{\sum_{k=1}^n (x_k - \bar{x}_n)^2},$$

$$\hat{\beta}_{0c} = \frac{1}{n} \sum_{k=1}^n y_k^+ - \hat{\beta}_{1c} \bar{x}_n,$$

$$\hat{\sigma}_c^2 = \frac{1}{n-2} \sum_{k=1}^n (y_k^+ - \hat{y}_k^+)^2,$$

where $\bar{x}_n = \frac{1}{n} \sum_{k=1}^n x_k$, $\hat{y}_k^+ = \hat{\beta}_{0c} + \hat{\beta}_{1c} x_k$, and

$$y_k^+ = \begin{cases} y_k & , k = 1, ..., m \\ \hat{y}_k^+ & , k = m + 1, ..., n. \end{cases}$$

Step 7. Compute stochastic mean $\hat{\mu} = \hat{\beta}_{0c} + \hat{\beta}_{1c}\bar{x}_n + e$ where e is a residual drawn randomly from $N(0, \hat{\sigma}_c^2)$.

Step 8. Repeat steps (1)-(7) M times to form the M multiple data sets.

Chapter 6

Simulation Methods and Summary Results

Simulations were used to study how to compensate for nonresponse in sample surveys. There were three compensation methods used: nonrespondent subsampling, weighting adjustment procedures and imputation methods. Three major criteria were used to investigate the survey design and compensation methods when there were unit nonresponses: bias, variance and design effect. Section 6.1 describes the simulation methods. General results are presented in section 6.2. Summary results for surveys with full response are described in section 6.3. Section 6.4 shows summary results with ignored nonrespondents. Nonrespondent subsampling, weighting adjustment procedures and imputation methods are discussed in section 6.5, 6.6 and 6.7 respectively.

6.1 Simulation Methods

The population was simulated from the income data of the 1997 Thailand summary industrial survey report. The population was divided into 12 strata. Stratification

was on the basis of four levels of business size and three levels of business type. These levels of business size were person engaged between 10-19, 20-49, 50-99 and 100 or over. The business types were used with Thailand Industrial Code for Major Division: 31, 32 and 33. A normal distribution was used within each of the 12 strata to generate the Y variable of the population. The parameters for the normal distribution within each stratum were based on the income data from the 1997 Thailand summary industrial survey report. The noramlity assumption is not essential to the conclusion in nonresponse problems as long as sample sizes are large enough.

The population size was $N = 2,055$. Strata sizes ranged from $N_{11} = 180$ to $N_{34} = 169$. An auxiliary variable, X, was generated for each Y. The auxiliary variable was correlated with the study variable. An example of an auxiliary variable for the Thailand industrial survey is cost of production when income is the study variable.

Three different sample sizes were used, $n = 15\%$, 30% and 50% of population size. Ten basic one-stage survey designs: *SRSWR*, *SRSWOR*, *STWR*, *STWOR*, *PTWR*, *PTWOR*, *USRSWR*, *USRSWOR*, *USTWR* and *USTWOR*[1], were used for comparison of efficiencies.

Sampling was with equal and unequal probability of selection. The auxiliary variable was used for unequal probability sampling. There were two different correlation coefficients between the study variable and auxiliary information: low correlation and high correlation. The high correlation coefficients were approximately between 0.7-0.9 and the low correlation coefficients were approximately between 0.1-0.3. The auxiliary variable, X, was generated from a normal distribution that was based on the distribution of the Y variable. The distribution for the auxiliary variable was such that X had the desired correlation with Y (*Dagpunar*, 1989).

The simulated survey either had full response or some nonresponse. There were

[1]See appendix for notation

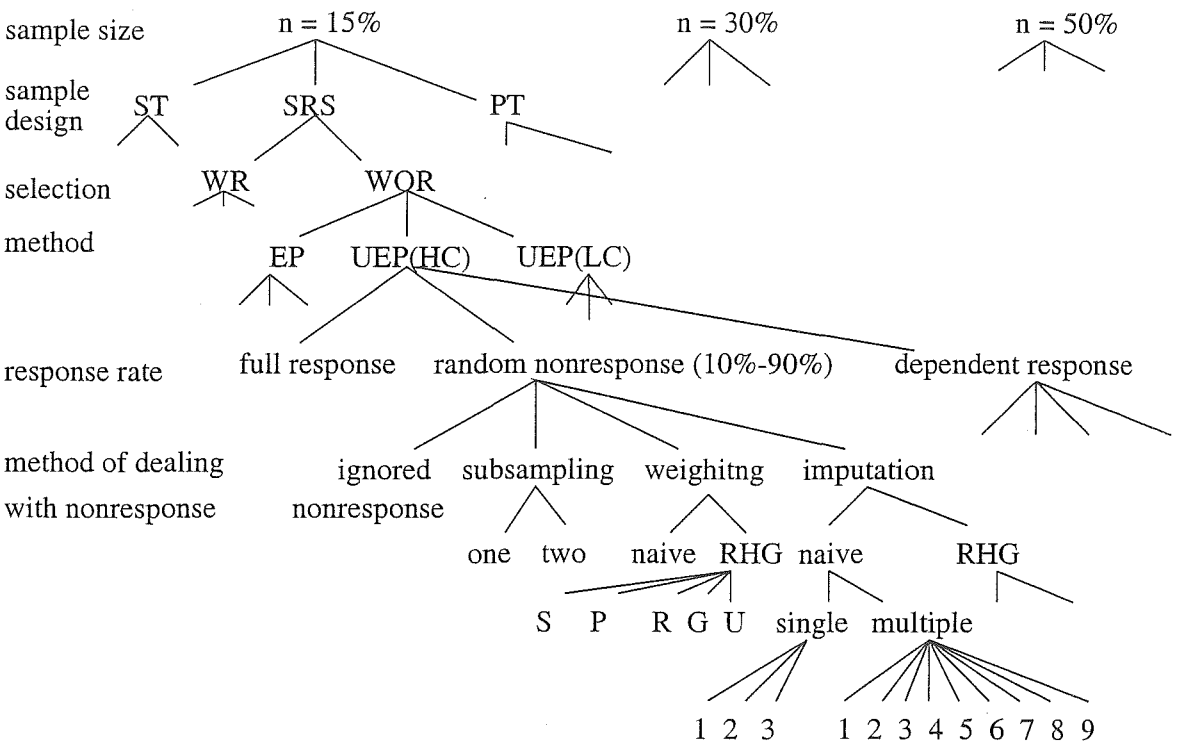


Figure 6.1: Flow Chart of Methodology

Note: Incomplete branches mean the tree is replicated at that node. Notations 1 to 3 in single and multiple imputation are random, sequential and stochastic regression methods respectively and notations 4 to 9 in multiple imputation are modified Wang's regression, approximated Bayesian bootstrap, fully normal, adjusted fully normal, mixture model with follow-up data and mixture model without follow-up data respectively. Post-stratified random sampling is used with equal probability sampling. *RHG* models are used in *SRS*. There are generally four *RHG* models: sample-based, population-based, raking ratio and general-based methods denoted by S, P, R and G respectively. bias-removal method is the fifth *RHG* model used in unequal probability sampling with replacement. There are 4,572 simulation study cases.

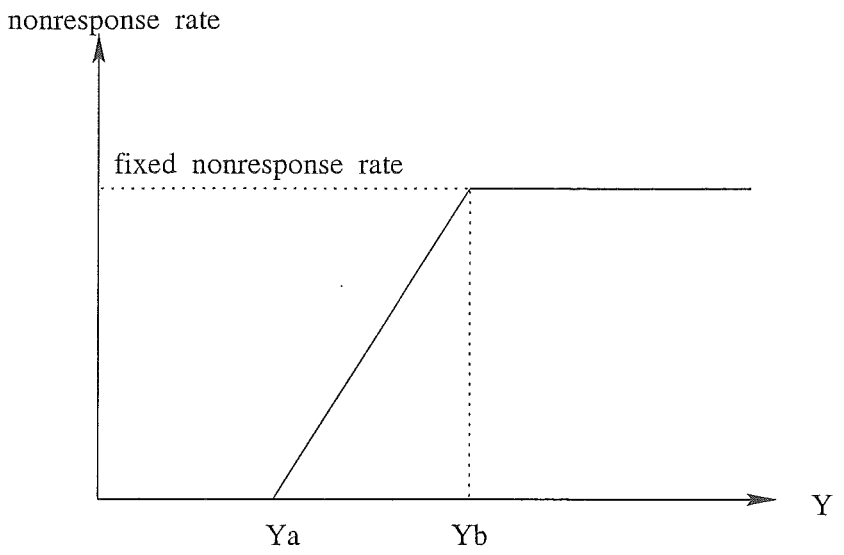


Figure 6.2: Dependent Nonresponse Pattern

two patterns for nonresponse: random and dependent response. The model for random response was a uniform distribution. Response rates were: 10%, 30%, 50%, 70% and 90%. In the dependent response mechanism model for the study variable Y the probability of nonresponse equalled zero if the value of Y was less than Y_a. The probability of nonresponse equalled 0.7 if the value of Y was greater than Y_b and it was a linear function between 0 to 0.7 if the value of Y was between Y_a and Y_b.

When there was nonresponse it was either ignored or compensated for. There were three methods used for compensating the unit nonresponse: nonrespondent subsampling, weighting adjustment and imputation methods. The two nonrespondent subsampling schemes were one-subsampling and two-subsampling. For weighting adjustment procedures two response models were used: naive and *RHG* model. Imputation methods are alternative methods to use to compensate missing data. There were four general schemes used: the naive or *RHG* model in single imputations and multiple imputations. $M = 3$ was studied in multiple imputation methods.

Not all of these methods were used for all sample survey designs proposed. Thus, the experimental design used for simulations is not balanced. It may have nevertheless been plausible to have used the experimental design structure of the simulations to find the most important factors in controlling nonresponse. For example, sequential imputation was not used in unequal probability sampling, and random imputation was not used with some random response cases for without replacement. For mixture model without follow-up data methods were divided into four models based on the difference among the estimated regression coefficients by using the coefficient of variation in β_0 and β_1 (C_β and C_η respectively): model A ($C_\beta = 0$, $C_\eta = 0$), model B ($C_\beta = 0$, $C_\eta = 0.02$), model C ($C_\beta = 0.02$, $C_\eta = 0$) and model D ($C_\beta = 0.02$, $C_\eta = 0.02$) (More details are in section 5.3.4.1).

Each survey design was simulated 1,000 times. The sample mean $\hat{\mu}$ and square root of the estimated variance of $\hat{\mu}$ were computed in each simulation. These statistics were then averaged to give $\bar{\hat{\mu}}$ and $\bar{s}_{\hat{\mu}}$. The bias was computed as the difference between the true population mean and the average of sample mean, $\bar{\hat{\mu}} - \mu$. The relative bias was $\frac{\bar{\hat{\mu}} - \mu}{\mu}$. The coefficient of variation of the estimator was computed as $CV = \frac{\bar{s}_{\hat{\mu}}}{\hat{\mu}}$. The design effect was calculated as $deff = \frac{\bar{s}^d_{\hat{\mu}}}{\bar{s}^{srs}_{\hat{\mu}}}$, where "d" means sampling design.

A diagram of the simulation pattern is given in figure 6.1. A list of all notation and major symbol used in this thesis is in appendix E.

6.2 General Results

There were same general trends in the results consistent with all designs. These were

- Sampling with replacement had a higher CV than sampling without replacement.

- CV decreased with increasing sample size, and as the response rate increased.
- Stratified random sampling and post-stratified random sampling had a lower CV than simple random sampling.
- CV was lower for sampling with unequal probability of selection compared with equal probability of selection.
- In unequal probability sampling, CV for the survey design with the high correlation was lower than with the low correlation.
- Stratified and post-stratified random sampling had the smallest design effects.

The following sections discuss specific results for when there was full response and some nonresponse. Complete results are in appendices.

6.3 Full Response

Simulation results are summarised in three aspects: relative bias, CV and design effect.

i) All sampling designs[2] were unbiased. The relative bias was at most 0.74% in simple random sampling with or without replacement for unequal probability of selection with the low level of correlation between the study variable Y and the auxiliary variable. More details of the relative bias are in table D.1-D.6, D.13-D.20 and D.29-D.31.

ii) The CV in all sampling designs had the same trends as those described in section 6.2. The CV and relative bias are summarised in table 6.1. More details of the CV are in table D.7-D.12, D.21-D.28 and D.32-D.34.

[2] *SRSWR*, *SRSWOR*, *STWR*, *STWOR*, *PTWR*, *PTWOR*, *USRSWR*, *USRSWOR*, *USTWR* and *USRSWOR*

Table 6.1: Relative bias and CV for full response with sampling designs varying by sample size (n = 15%, 30%, 50%)

Sample-selection	Relative Bias	CV		
		15%	30%	50%
srswr	0.0031	5.7400	4.0600	3.1400
srswor	0.0031	5.3000	3.3800	2.2100
stwr	0.0000	0.0014	0.0010	0.0008
stwor	0.0000	0.0013	0.0008	0.0005
ptwr	0.0000	0.0015	0.0010	0.0008
ptwor	0.0000	0.0014	0.0009	0.0006
usrswr:high	0.0064	2.9500	1.9900	1.6400
usrswor:high	0.0064	2.6700	1.7600	1.2100
ustwr:high	0.0000	0.0012	0.0008	0.0006
ustwor:high	0.0000	0.0011	0.0007	0.0005
usrswr:low	0.0074	2.9500	2.0500	1.6500
usrswor:low	0.0074	2.6700	1.7600	1.2100
ustwr:low	0.0000	0.0041	0.0029	0.0022
ustwor:low	0.0000	0.0039	0.0028	0.0022

iii) In general the sampling designs with the smallest *deff* were stratified random sampling, post-stratified random sampling and unequal probability of selection for stratified random sampling. Simple random sampling with unequal probability of selection had a smaller *deff* than with equal probability of selection. The *deffs* are summarised in table 6.2.

Rank order of designs are given in table 6.3. The design with the smallest *deff* in the list was stratified sampling with unequal probability of selection and where the correlation with the auxiliary variable was high. This was

Table 6.2: Design Effect for full response for different sample designs and sample sizes. Designs with a *deff* less than 1 are considered more powerful than *SRS*

Sampling Design	15wr	15wor	30wr	30wor	50wr	50wor
srs	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
st	6.1072e-8	6.1228e-8	5.9767e-8	6.1371e-8	8.2137e-8	6.0614e-8
pt	6.3683e-8	6.3581e-8	6.2050e-8	6.2643e-8	6.0831e-8	6.1549e-8
husrs	0.2414	0.2235	0.2422	0.2058	0.2411	0.2441
hust	4.8089e-7	4.3057e-8	4.9593e-8	6.0309e-8	4.8988e-7	4.7265e-8
lusrs	0.2418	0.2209	0.2372	0.2182	0.2627	0.2294
lust	4.8739e-7	5.2826e-7	5.2217e-7	6.3723e-7	5.3327e-7	9.6244e-7

Note: Notations "srs" = simple equal probability random sampling, "st" = stratified equal probability random sampling, " pt" = post-stratified equal probability random sampling, "husrs" = simple unequal probability random sampling with the high correlated case, "hust" = stratified unequal probability random sampling with the high correlated case, "lusrs" = simple unequal probability random sampling with the low correlated case, "lust" = stratified unequal probability random sampling with the low correlated case.

Table 6.3: Ascending order in Design Effect for full response. Rank 1 is the most powerful design, ie smallest *deff*

Sampling Design	srs	st	pt	husrs	hust	lusrs	lust
rank	7	2	3	5	1	6	4

considered the most powerful design.

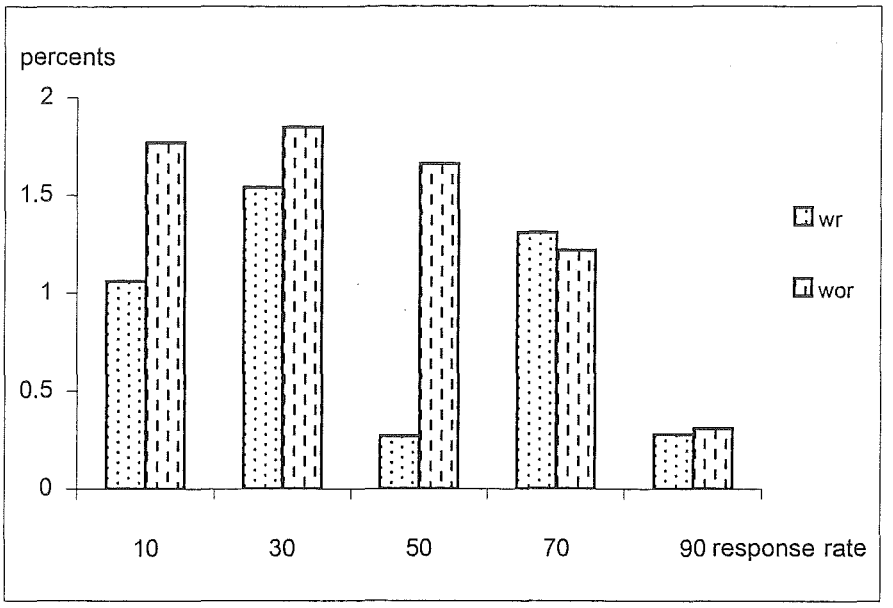


Figure 6.3: Relative bias for ignored nonrespondents in *SRSWR* and *SRSWOR* with 30% sample size and varying response rate

6.4 Ignored Nonrespondents

Three major conclusions were

i) When there was nonresponse, and it was ignored, the simple random sampling gave biased results (fig 6.3). The bias was considerably worse when the nonresponse was dependent on the study variable Y. For example, the absolute relative biases for *SRSWOR* and 15% sample size was at most 5.03% (table D.1) for the random response mechanism and was approximately 36.51% (table D.29) for the dependent response mechanism. **Stratified and post-stratified random sampling had negligible sampling bias** (table D.1-D.6, D.13-D.20 and D.29-D.31). However, clearly stratification is critical for sampling efficiency.

ii) The CV was higher with nonresponse compared with full response. The CV

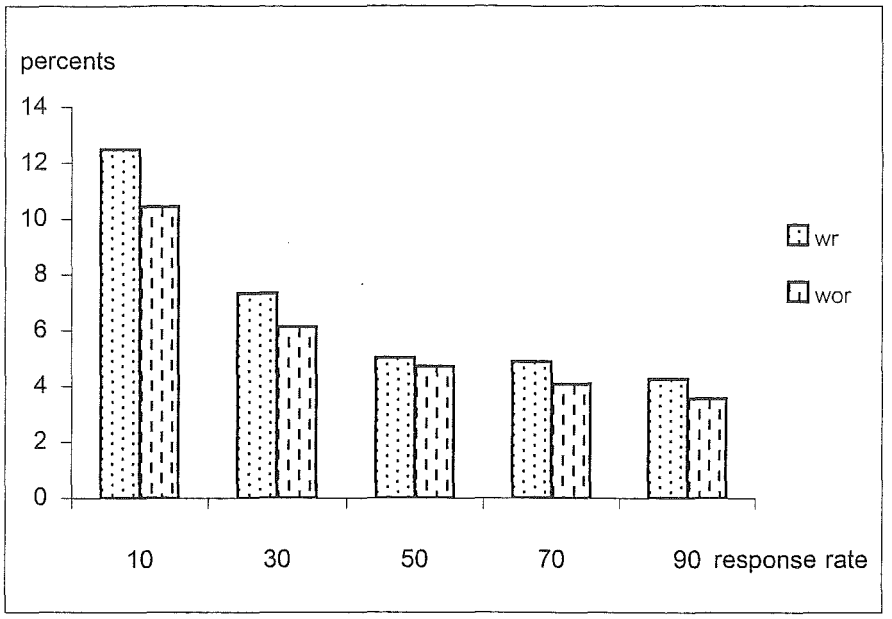


Figure 6.4: CV for ignored nonrespondents in *SRSWR* and *SRSWOR* with 30% sample size and varying response rate

was high when there was a high level of nonresponse. For example, in *SRSWR* with 30 % sample size the CVs were 12.49%, 7.33%, 5.62%, 4.89% and 4.26 % for the 10%, 30%, 50%, 70% and 90% of level of random response (fig 6.4).

The CV in *SRSWR* and *SRSWOR* for the full response and ignored nonrespondents are summarised in table 6.4. More details of the CV are in table D.7-D.12, D.21-D.28 and D.32-D.34.

iii) The sampling design with the smallest *deff* was stratified and post-stratified random sampling for both the random and the dependent response mechanism. In simple random sampling for both the random and the dependent response mechanism the *deffs* were larger than 1 for equal probability sampling and some cases for unequal probability sampling (table E.1-E.8). Rank order of designs when nonrespondents were ignored are given in table 6.5 for sampling with and without replacement.

Table 6.4: CV for ignored nonrespondent with five levels of random response and one level of dependent response in simple random sampling with or without replacement varying by sample size (n = 15%, 30%, 50%)

Sample-selection	response rate					
	10%	30%	50%	70%	90%	*dependent*[1]
15-wr-full	5.73	5.74	5.74	5.74	5.74	5.74
15-wr-ignored	18.23	10.37	8.24	6.97	6.08	6.84
30-wr-full	4.05	4.05	4.04	4.04	4.04	4.04
30-wr-ignored	12.49	7.33	5.62	4.89	4.26	5.09
50-wr-full	3.14	3.14	3.14	3.14	3.14	3.14
50-wr-ignored	9.75	6.02	4.58	3.74	3.33	3.83
15-wor-full	5.29	5.29	5.29	5.29	5.30	5.31
15-wor-ignored	16.84	9.54	7.61	6.43	5.62	6.47
30-wor-full	3.38	3.38	3.38	3.38	3.38	3.38
30-wor-ignored	10.45	6.13	4.71	4.09	3.57	4.62
50-wor-full	2.21	2.21	2.21	2.21	2.21	2.1
50-wor-ignored	6.83	4.52	3.23	2.64	2.35	3.20

Note: This table is summarised from table D.7, D.8 and D.32. 1 means nonresponse rate up to 70%.

6.5 Nonrespondent Subsampling

The summary results for nonrespondent subsampling methods were:

i) The relative bias from nonresponse was reduced with nonrespondent subsampling. For example, in *SRSWOR* for 15% sample size the relative bias was at most 0.43% with the 70% response rate in two-subsampling scheme (fig 6.5). More details for relative bias of nonrespondent subsampling with equal probability sampling design are in table D.35-D.37. In general, use

Table 6.5: Ascending order in Design Effect for ignored nonrespondent in sampling with and without replacement for the average of random response mechanism.

Sampling Design	15wr	15wor	30wr	30wor	50wr	50wor
srs	7	7	7	7	7	7
st	2	2	1	2	2	2
pt	3	4	3	4	3	3
husrs	5	5	6	6	5	5
hst	1	1	2	1	1	1
lusrs	6	6	5	5	6	6
lust	4	3	4	3	4	4

Note: Ascending order is computed from the average of five level of random response solutions.

Table 6.6: Relative bias in nonrespondent one-subsampling and two-subsampling scheme for 50% random response rate for simple random sampling with and without replacement varying with sample size (n = 15%, 30%, 50%)

Sampling Design	15srswr	15srswor	30srswr	30srswor	50srswr	50srswor
one-subsampling	0.1015	0.0746	0.2693	0.0787	0.0404	0.3428
two-subsampling	0.0269	0.0476	0.0342	0.2061	0.2942	0.3832

of the one-subsampling scheme gave a smaller relative bias than the two-subsampling scheme. For example, over half (18) of the 30 sample design combinations in equal probability of selection for SRS[3] for the one-subsampling scheme had a smaller relative bias than for the two-subsampling scheme. **For the dependent response mechanism in *SRS* all cases with the one-subsampling scheme had a smaller relative bias than with two-**

[3]two types of selection procedure (WR and WOR), three level of sample size and five levels of response rate: the number of sample designs = 2×3×5

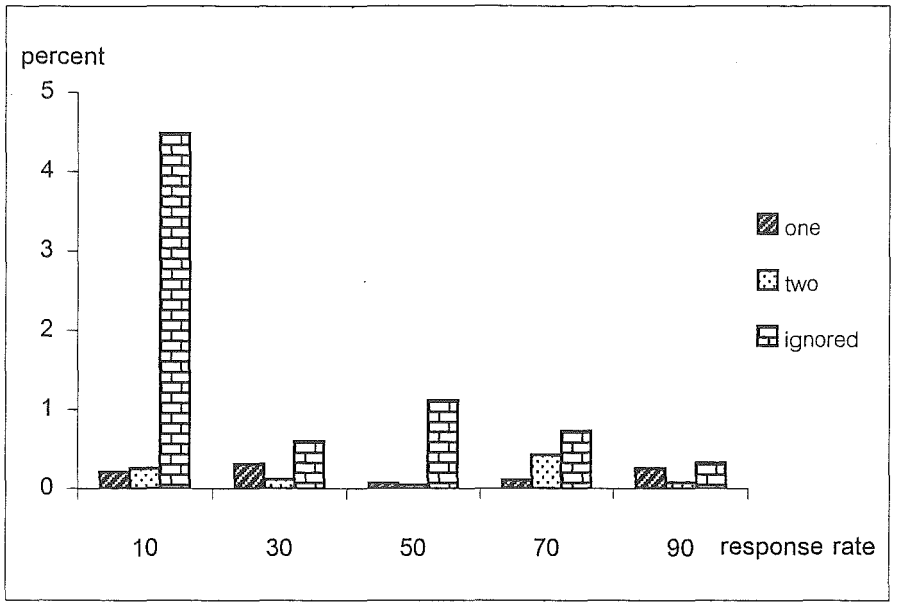


Figure 6.5: Relative bias for one and two subsampling scheme in *SRSWOR* with 15% sample size and varying response rate

subsampling scheme (table D.53). Details of the bias for nonrespondent subsampling are in table D.35-D.37, D.41-D.46 and D.53-D.55.

Sampling with and without replacement in nonrespondent subsampling did not affect the bias. The sample size also did not affect the size of the bias (table 6.6).

ii) The CV for nonrespondent subsampling was reduced compared with when nonrespondents were ignored. Two-subsampling scheme had a larger variance than one-subsampling scheme. However, the difference variance between one-subsampling and two-subsampling scheme were small when the response rate was high. For example, the CV in *SRSWOR* with the 50% random response rate and 15% sample size was reduced from 7.61% for ignored nonrespondents to 7.00% and 7.02% for one-subsampling and two-subsampling scheme respectively (fig 6.6). The amount of CV reduction in *SRSWR* was approximately

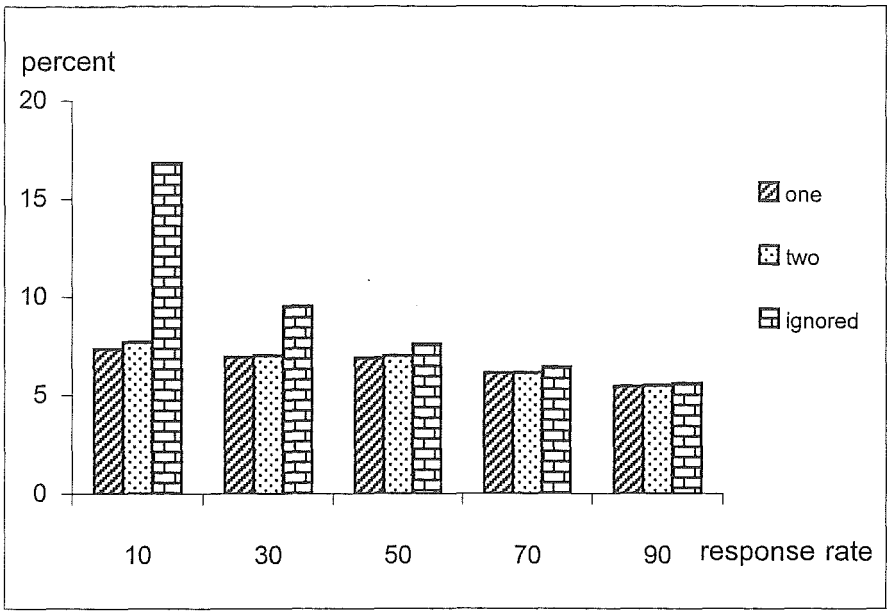


Figure 6.6: CV for one and two subsampling scheme in *SRSWOR* with 15% sample size and varying response rate

1% – 56% for one-subsampling scheme and 0.7% – 54% for two-subsampling scheme with the level of random response rate from 90% to 10%. The CV for simple random sampling and nonrespondent subsampling with random response mechanism was reduced with increasing response rates (table 6.7).

iii) The sampling design for nonrespondent subsampling with the smallest *deff* was stratified and post-stratified random sampling for both the random and the dependent response mechanism (table 6.8). In simple random sampling the *deffs* for both the random and the dependent response mechanism were larger than 1 for equal probability sampling and some cases for unequal probability sampling with both the high and low level of correlation cases (table E.1-E.8).

Table 6.7: CV for one and two nonrespondent subsampling with five levels of random response and one level of dependent response in simple random sampling both with or without replacement varying by sample size (n = 15%, 30%, 50%)

Sample-selection	response rate					
	10%	30%	50%	70%	90%	*dependent*[1]
15-wr-one	8.06	7.49	7.06	6.66	5.99	6.11
30-wr-one	5.50	5.27	4.91	4.59	4.22	4.47
50-wr-one	4.28	4.18	3.87	3.57	3.29	3.41
15-wor-one	7.34	6.93	6.99	6.15	5.47	5.53
30-wor-one	5.47	5.71	4.63	3.87	3.38	4.31
50-wor-one	3.38	3.06	2.60	2.54	2.30	3.17
15-wr-two	8.34	7.64	7.06	6.69	6.01	6.78
30-wr-two	5.71	5.46	5.03	4.61	4.22	4.92
50-wr-two	5.12	4.91	4.45	3.68	3.30	3.81
15-wor-two	7.71	7.03	7.01	6.15	5.52	5.78
30-wor-two	5.57	5.73	4.65	3.98	3.39	4.31
50-wor-two	4.10	3.08	2.69	2.58	2.32	3.17

Note: Notations for sample-selection 15, 30, 50, wr, wor, one and two are 15%, 30% and 50% of sample size, sampling with replacement, sampling without replacement, one-subsampling scheme and two-subsampling scheme respectively. This table is summarised from table D.38-D.40 and D.56. 1 means nonresponse rate up to 70%.

6.6 Weighting Adjustment Methods

The summary results for weighting adjustment methods for the naive and *RHG* models were:

i) The relative bias from nonresponse was reduced, compared with ignoring non-response, with three *RHG* models in the weighting adjustment method in equal

Table 6.8: Ascending order in Design effect for nonrespondents subsampling with random and dependent response mechanism with varying sample sizes (n = 15%, 30%, 50%)

Sampling Design	15wr	15wor	30wr	30wor	50wr	50wor	depwr	depwor
srs-one	15	15	15	15	15	15	15	15
srs-two	16	16	16	16	16	16	16	16
st-one	1	1	1	1	1	1	1	3
st-two	2	3	3	2	2	3	2	5
pt-one(pt,st)	3	5	2	3	3	4	4	4
pt-one(pt,pt)	4	6	4	7	4	5	6	6
pt-two(pt,st,st)	5	7	5	5	5	6	8	7
pt-two(pt,pt,pt)	6	8	6	8	6	8	7	8
husrs-one	11	12	11	11	11	11	11	11
husrs-two	13	13	13	12	12	14	13	14
hust-one	7	2	7	6	7	2	3	1
hust-two	9	4	9	4	9	7	5	2
lusrs-one	12	11	12	13	13	12	12	12
lusrs-two	14	14	14	14	14	13	14	13
lust-one	8	9	8	9	8	9	9	9
lust-two	10	10	10	10	10	10	10	10

Note: depwr and depwor: nonresponse rate up to 70%.

probability of selection for simple random sampling with random response. The three models were **sample-based, population-based and raking ratio** methods. For example, with a 50% random response rate and 30% sample size with *SRSWR* the relative bias for ignored nonrespondents was 0.27%. This reduced to 0.22%, 0.00% and 0.04% for sample-based, population-based and raking ratio methods respectively (table 6.9). The population-based adjustment method was unbiased. **If the sample size was large enough,**

Table 6.9: Relative bias for the naive and *RHG* models with 30% sample size in *SRSWOR* varying by levels of random response (*rate* = 10%, 30%, 50%, 70%, 90%)

weighting method	response rate				
	10%	30%	50%	70%	90%
ignored	1.0617	1.5412	0.2735	1.3061	1.3061
naive	1.0617	1.5412	0.2735	1.3061	1.3061
rhg:g	1.0617	1.5412	0.2735	1.3061	1.3061
rhg:s	0.3687	0.1139	0.2206	0.0932	0.1441
rhg:p	0.0010	0.0000	0.0000	0.0000	0.0000
rhg:r	0.1647	0.0383	0.0383	0.0383	0.0383

Note: Notations: "ignored" = ignored nonrespondents, "naive" = naive method, "rhg:g" = general-based method, "rhg.s" = sample-based method, "rhg.p" = population-based method and "rhg.r" = raking ratio method.

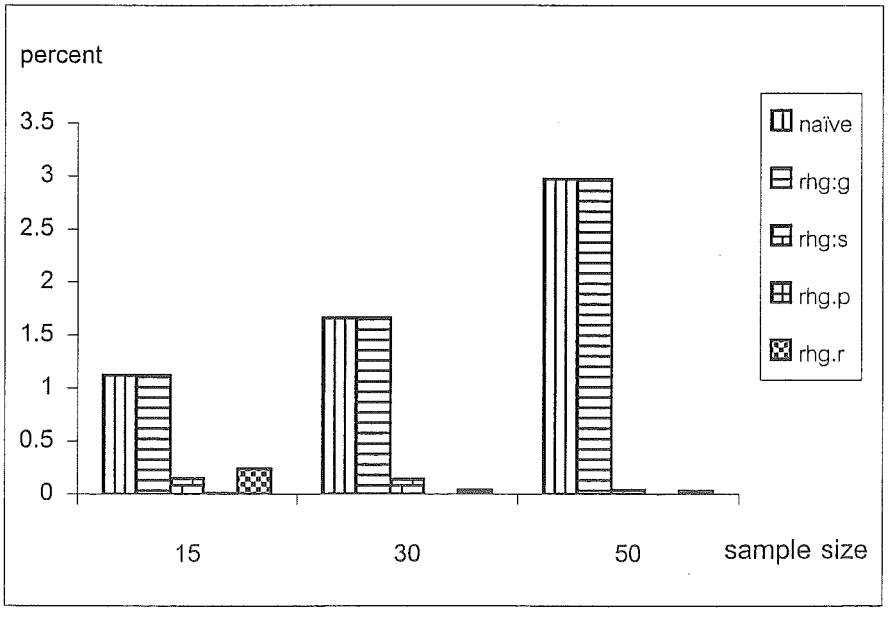


Figure 6.7: Relative bias on weighting adjustment procedure with naive and *RHG* models in srswor with 50% level of random response rate and varying sample size

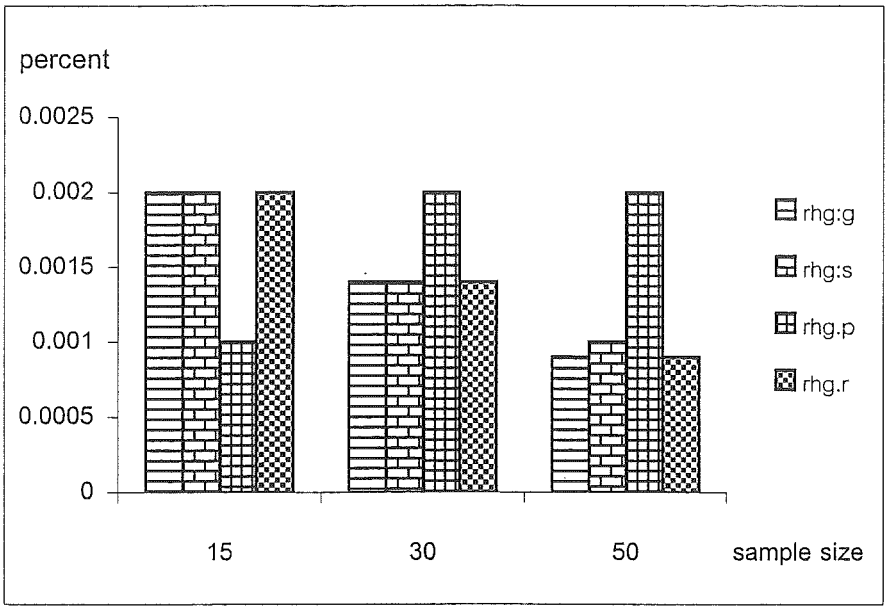


Figure 6.8: CV on weighting adjustment procedure with *RHG* models in srswor with 50% level of random response rate and varying sample size

raking ratio was also unbiased. However, the general-based method and naive model did not reduce the bias (fig 6.7). For the dependent response mechanism with *SRSWR* and *SRSWOR* three *RHG* models, sample-based, population-based and raking ratio methods successfully reduced the bias (table 6.10, details are in table D.89).

For unequal probability sampling with the random response mechanism both for the high and low level of correlation the relative bias in *USRSWR* was reduced only with the bias-removal method and in *USRSWOR* was reduced only with the sample-based method. However in some cases the general-based, population-based and raking ratio methods in *USRSWOR* reduced the relative bias (table D.77-D.82). For the dependent response mechanism for unequal probability sampling both with the high and low level of correlation the relative bias in *USRSWR* was reduced only with the bias-removal method and in *USRSWOR* was reduced with sample-based, population-based and raking ratio methods (table D.90-D.91). If weighting methods in *USRSWR* were rescaled to give an unbiased estimator, the results in relative bias is the same as in a

Table 6.10: Relative bias for the naive and *RHG* models with dependent response mechanism in *SRSWOR* varying by levels of sample size (n = 15%, 30%, 50%)

weighting method	sample size		
	15%	30%	50%
ignored	36.3907	32.4184	33.4107
naive	36.3907	32.4184	33.4107
rhg:g	36.3907	32.4184	33.4107
rhg:s	0.7106	0.0725	0.0787
rhg:p	0.0124	0.0033	0.0010
rhg:r	0.2382	0.0383	0.0249

Note: Notations: "ignored" = ignored nonrespondents, "naive" = naive method, "rhg:g" = general-based method, "rhg.s" = sample-based method, "rhg.p" = population-based method and "rhg.r" = raking ratio method. Nonresponse rate is up to 70%.

bias-removal method. The relative bias with the high level of correlation in *USRS* was summarised in table 6.11. These results are included simply for completeness and to illustrate theorems 4.14, 4.18, 4.22 and 4.26.

ii) **The CV in weighting adjustment method was reduced with the *RHG* models compared with the ignored nonrespondents and the naive model** (fig 6.8 and table 6.12). For equal probability sampling design the CV of the population-based method in *SRS* was the same to the CV of the naive model in post-stratified random sampling (see table 6.4 and 6.12).

iii) The sampling design with the smallest *deff* was stratified and post-stratified random sampling for both the random and the dependent response mechanism. In some situations simple random sampling gave the smallest *deff*; the population-based method for equal probability sampling, the bias-removal method for unequal probability sampling with replacement and the sample-

based method for unequal probability sampling without replacement. However, with equal probability of selection for simple random sampling with both the random and the dependent response mechanism the *deffs* were smaller than 1 for sample-based and raking ratio methods and for some cases with general-based method. For unequal probability sampling without replacement in some cases with random and dependent response mechanism the *deff* was smaller than 1 with population-based and raking ratio methods. Rank order of *deffs* are given in table 6.13. More details on the *deffs* are in table D.95-D.106.

6.7 Imputation Methods

Full results for imputation methods are given in table D.107-D.262. The summary results for single and multiple imputation methods for the naive and *RHG* models were:

i) RELATIVE BIAS:

The relative bias with the naive model was reduced if stochastic regression, modified Wang's regression, mixture model with follow-up data and mixture model without follow-up data methods had been chosen for compensating the missing data (fig 6.9). The relative bias was also reduced if the three single imputation and the nine multiple methods were used with the *RHG* model in sampling design. Compared with ignored nonrespondents, the relative bias was reduced by imputation methods and results are summarised in table 6.14. More details on relative bias with imputation methods are in table D.107-D.112, D.125-D.130, D.179-D.184, D.197-D.202, D.215-D.216, D.221-224, D.239-D.242 and D.251-D.254.

ii) CV in SINGLE IMPUTATION:

ii.1) Equal probability sampling

Table 6.11: Relative bias for the naive and *RHG* models with dependent response mechanism in *USRSWR* and *USRSWOR* varying by levels of sample size (n = 15%, 30%, 50%)

weighting method	sample size		
	15%	30%	50%
USRSWR			
ignored	1.8220	7.8161	4.5223
naive	1.8220	7.8161	4.5223
rhg:g^1	94.3742	93.9902	94.0991
rhg:s^1	90.3963	91.2990	91.2210
rhg:p^1	90.4095	91.2985	91.2089
rhg:r^1	91.4223	91.3018	91.2117
rhg:u	0.0010	0.0000	0.0000
USRSWOR			
ignored	4.8434	7.9093	7.3811
naive	4.8434	7.9093	7.3811
rhg:g	36.3938	32.3852	33.5288
rhg:s	0.0000	0.0000	0.0000
rhg:p	0.2299	0.2227	0.4796
rhg:r	0.0269	0.1833	0.4547

Note: Notations: "ignored" = ignored nonrespondents, "naive" = naive method, "rhg:g" = general-based method, "rhg.s" = sample-based method, "rhg.p" = population-based method, "rhg.r" = raking ratio method and "rhg.u" = bias-removal method. "1" indicates that if estimator in this method is rescaled to get an unbiased estimator, the relative bias is the same as in a bias-removal method. Nonresponse rate is up to 70%.

For the random and dependent response mechanism with the naive model the variance in stochastic regression was the smallest followed by random and se-

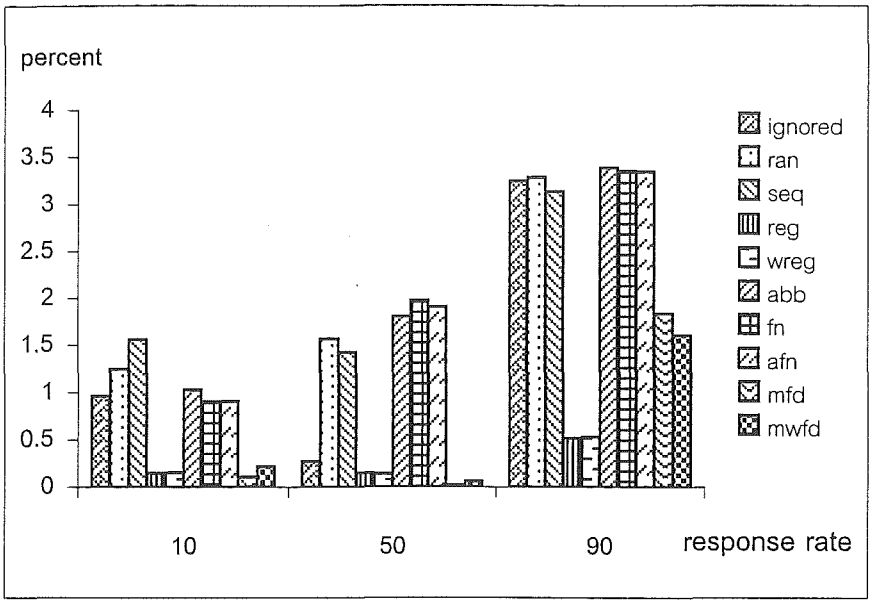


Figure 6.9: Relative bias on imputation method with the naive model in 15% sample size of *SRSWR* varying level of random response rate

quential methods respectively except in the case of 10% of sample size. However, if the *RHG* model was used with single imputation methods, the CV was even smaller and was at most 0.04%. The CV for the *RHG* model had the same pattern as for the naive model. More details about the CV are in table D.113-D.118 and D.225-D.226.

Both with random response mechanism and dependent response mechanism, when there was high correlation between the auxiliary variable and study variable the CV for stochastic regression method was smaller than for the low correlation case (table D.167-D.172, D.185-D.190 and D.235-D.236).

ii.2) Unequal probability sampling

With a naive model the variance in stochastic regression method was the smallest for the random and dependent response mechanism followed by random imputation method. When the *RHG* model was used with single imputation

methods, the CV was decreased further and was at most 0.0036%. The CV for the *RHG* model had the same pattern as for the naive model. More details about the CV are in table D.203-D.208 and D.255-D.258.

Both with random response mechanism and dependent response mechanism, when there was high correlation between the auxiliary variable and study variable the CV for stochastic regression method was smaller than for the low correlation case (table D.203-D.208, D.217-D.218 and D.255-D.258).

iii) CV in MULTIPLE IMPUTATION:

iii.1) Equal probability sampling

For the random and dependent response mechanism with the naive model there was little variance difference in variuos imputation methods on the high level of response rate. The variance in modified Wang's regression was the smallest followed by stochastic regression, mixture model without follow-up data, mixture model with follow-up data, adjusted fully normal, approximate Bayesian bootstrap, fully normal, random and sequential methods respectively except in the case of 10% level of random response in 10% of sample size where stochastic regression, modified Wang's regression, mixture model with follow-up data and mixture model without follow-up data methods gave a larger variance (fig 6.10 and table D.131-D.136).

Comparing the CV for mixture models with follow-up and without follow-up data methods, model A had the smallest CV followed by model C, mixture model with follow-up data, model B, and model D respectively (table D.185-D.190). However, if the *RHG* model were used with multiple imputation methods, the CV was decreased further and was at most 0.14%. The CV for the *RHG* model had the same pattern as for the naive model. The CV in the same method for single imputation was smaller than for multiple imputation.

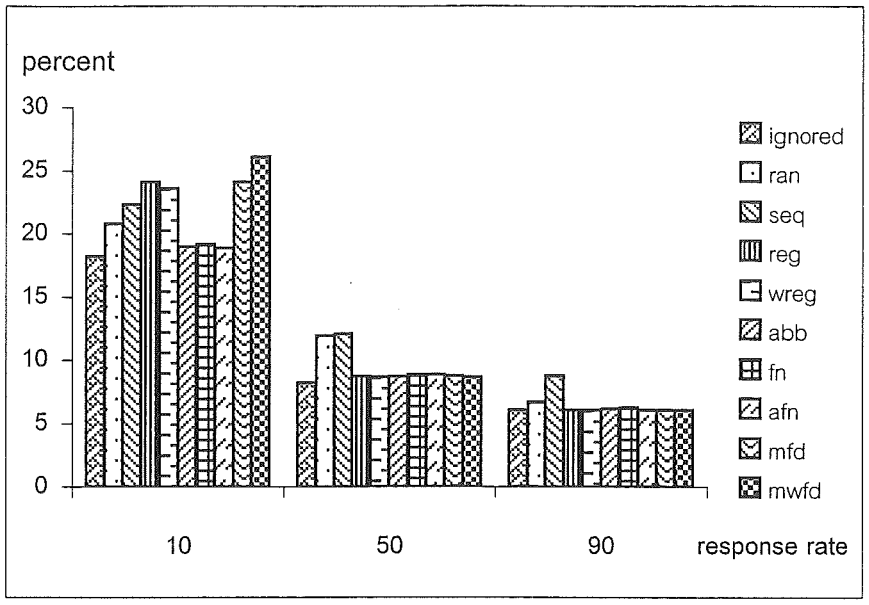


Figure 6.10: CV on imputation method with the naive model in 15% sample size of *SRSWR* and varying level of random response rate

Both with random response mechanism and dependent response mechanism, when there was high correlation between the auxiliary variable and study variable the CV for stochastic regression, modified Wang's regression, mixture model with follow-up data and mixture model without follow-up data methods was smaller than for the low correlation case (table D.167-D.172, D.185-D.190, D.235-D.236 and D.245-D.246).

iii.2) Unequal probability sampling

As with the single imputation for random response mechanism and dependent response mechanism, the CV for multiple imputation had the same pattern and was larger than for single imputation (table D.203-D.208, D.217-D.218 and D.255-D.258).

iv) The sampling design with the smallest *deff* was stratified and post-stratified random sampling for both the random and the dependent response mechanism. In simple random sampling with the *RHG* model also gave the smallest *deff*. The efficient sampling design for simple, stratified and post-stratified random

sampling had the same pattern both for single and multiple imputation. Single imputation methods such as random, sequential and stochastic regression had a lower *deff* than multiple imputation with the same method. However, in some multiple imputation methods such as approximated Bayesian bootstrap, fully normal, adjusted fully normal or methods related with regression procedures such as modified Wang's regression, mixture model with follow-up data and mixture model without follow-up data methods, the *deff* was lower than for single random or sequential methods with the high level of correlation between the auxiliary variable and study variable (table E.17-E.24).

Table 6.12: CV of weighting adjustment method with naive and *RHG* models for five levels of random response and one level of dependent response in simple random sampling both with and without replacement varying by sample size (n = 15%, 30%, 50%)

Sample-selection	response rate					
	10%	30%	50%	70%	90%	*dependent*[1]
15-wr-naive	19.0294	11.7863	10.0569	9.0742	8.3875	8.5029
15-wr-rhg:g	0.0050	0.0029	0.0025	0.0023	0.0021	0.0026
15-wr-rhg:s	0.0047	0.0031	0.0025	0.0023	0.0021	0.0034
15-wr-rhg:p	0.0045	0.0030	0.0025	0.0023	0.0021	0.0034
15-wr-rhg:r	0.0045	0.0030	0.0025	0.0023	0.0021	0.0033
15-wor-naive	16.8358	9.5418	7.6078	6.4293	5.6223	6.5427
15-wor-rhg:g	0.0049	0.0025	0.0020	0.0017	0.0014	0.0018
15-wor-rhg:s	0.0047	0.0027	0.0020	0.0017	0.0014	0.0030
15-wor-rhg:p	0.0044	0.0026	0.0020	0.0017	0.0014	0.0030
15-wor-rhg:r	0.0044	0.0026	0.0020	0.0017	0.0014	0.0029
30-wr-naive	13.1130	8.3597	6.9136	6.3781	5.8855	6.4346
30-wr-rhg:g	0.0033	0.0021	0.0018	0.0016	0.0015	0.0018
30-wr-rhg:s	0.0034	0.0022	0.0018	0.0016	0.0015	0.0026
30-wr-rhg:p	0.0034	0.0022	0.0018	0.0016	0.0015	0.0026
30-wr-rhg:r	0.0033	0.0022	0.0018	0.0016	0.0015	0.0026
30-wor-naive	10.4513	6.1334	4.7153	4.0862	3.5680	4.6282
30-wor-rhg:g	0.0032	0.0018	0.0014	0.0011	0.0009	0.0015
30-wor-rhg:s	0.0034	0.0018	0.0014	0.0011	0.0009	0.0021
30-wor-rhg:p	0.0033	0.0018	0.0014	0.0011	0.0009	0.0023
30-wor-rhg:r	0.0034	0.0018	0.0014	0.0011	0.0009	0.0023
50-wr-naive	10.1646	6.8780	5.6091	4.8769	4.5904	4.8451
50-wr-rhg:g	0.0026	0.0016	0.0013	0.0012	0.0011	0.0014
50-wr-rhg:s	0.0028	0.0017	0.0013	0.0012	0.0011	0.0019
50-wr-rhg:p	0.0028	0.0017	0.0013	0.0012	0.0011	0.0019
50-wr-rhg:r	0.0028	0.0017	0.0013	0.0012	0.0011	0.0019
50-wor-naive	4.2613	4.2613	3.2317	2.6567	2.3526	3.2124
50-wor-rhg:g	0.0025	0.0013	0.0009	0.0008	0.0006	0.0017
50-wor-rhg:s	0.0026	0.0014	0.0010	0.0008	0.0006	0.0017
50-wor-rhg:p	0.0026	0.0014	0.0009	0.0008	0.0006	0.0015
50-wor-rhg:r	0.0026	0.0014	0.0009	0.0008	0.0006	0.0015

Note: Notations for sample-selection 15, 30, 50, wr, wor, naive, *RHG*, g, s, p and r are 15%, 30% and 50% of sample size, sampling with replacement, sampling without replacement, naive model, random homogeneity group model, general-based, sample-based, population-based and raking ratio methods respectively. This table is summarised from table D.74-D.76 and D.92. 1 means nonresponse rate up to 70%.

Table 6.13: Design effect of weighting adjustment method with naive and *RHG* models for five levels of random response and one level of dependent response in ten survey designs varying by sample size (n = 15%, 30%, 50%)

Sample-selection	15wr	30wr	50wr	15wor	30wor	50wor	depwr	depwor
srs-naive	12	11	12	18	18	18	12	18
srs-rhg-g	9	9	9	11	11	14	11	16
srs-rhg-s	7	8	8	8	8	8	8	8
srs-rhg-p	3	3	3	3	3	3	4	3
srs-rhg-r	8	7	6	7	7	7	7	7
st-naive	2	2	12	2	2	2	1	2
husrs-naive	11	10	11	17	16	16	9	13
husrs-rhg-g	14	19	13	10	15	17	20	15
husrs-rhg-s	13	14	17	4	4	4	15	5
husrs-rhg-p	17	17	16	12	12	11	14	10
husrs-rhg-r	15	20	18	14	9	9	16	9
husrs-rhg-u	4	4	4	-	-	-	3	-
hust-naive	1	1	1	1	1	1	2	1
lusrs-naive	10	12	10	16	17	15	10	14
lusrs-rhg-g	18	13	14	9	14	13	19	17
lusrs-rhg-s	19	15	20	6	6	6	18	6
lusrs-rhg-p	20	16	15	13	13	12	17	12
lusrs-rhg-r	16	18	19	15	10	10	13	11
lusrs-rhg-u	5	6	7	-	-	-	6	-
lust-naive	6	5	5	5	5	5	5	4

Note: Notations for sample-selection 15, 30, 50, wr, wor, naive, *RHG*, g, s, p, r and u are 15%, 30% and 50% of sample size, sampling with replacement, sampling without replacement, naive model, random homogeneity group model, general-based, sample-based, population-based, raking ratio and bias-removal methods respectively. This table is summarised from table D.95-D.106. Post-stratified random sampling with a naive model is considered a population-based method in simple random sampling. Symbol "-" means no method was used. depwr and depwor: nonresponse rate is up to 70%.

Table 6.14: Bias reduction in single and multiple imputation methods with the naive and the *RHG* model

Sample-selection	random naive	random rhg	*dependent*[a] naive	*dependent*[a] rhg
single imputation				
random	2	1	2	1
sequential	2	1	2	1
regression	2	1	1	1
multiple imputation				
random	2	1	2	1
sequential	2	1	2	1
regression	2	1	1	1
Wang's regression	2	1	1	1
abb	2	1	2	1
fully normal	2	1	2	1
adjusted fully norm	2	1	2	1
mixture with data	2	1	1	1
mixture without data	2	1	1	1

Note: Symbol "1" and "2" mean reduce and not reduce the bias respectively. a means nonresponse rate up to 70%.

Chapter 7

Conclusions

In this chapter, conclusions from the unit nonresponse simulations with the ten basic survey designs combined with three compensation methods are discussed in section 7.1. Section 7.2 presents algorithm for dealing with nonresponse. Summary of answers to the research questions for this thesis is presented in section 7.3.

7.1 Conclusion

In conclusion,

1. If auxiliary information is available in the sampling frame, an efficient sampling design should be used such as stratified random sampling where there is qualitative auxiliary information or unequal probability sampling where there is quantitative auxiliary information. Sampling error can be decreased by the use of the auxiliary information for stratification and for unequal probability sampling. The quantitative auxiliary information used in unequal probability sampling should be highly correlated with the study variable. If the level of the correlation between the study variable and the auxiliary information is low, equal probability sampling design should be used. However, if the auxiliary information is not available in the

planning stage but can be obtained during the data collection phase, post-stratification is useful for increasing the precision of the estimator. The sample unit selection procedure of sampling without replacement can give more precision compared with sampling with replacement even if these two schemes have the same effect on bias reduction. These results are of course well known; see for example *Cochran* (1977).

2. When the survey is conducted it is inevitable that there will be some nonresponse. Nonrespondent subsampling is useful during the data collection phase and weighting adjustment procedures and imputation methods can be used during the estimation phase.

2.1) Nonrespondent Subsampling: This method can give an unbiased estimator with both random and dependent response mechanisms. One-subsampling schemes are more efficient than two-subsampling schemes. This is because the sampling variance at each subsample (e.g., equation 3.2) step is cumulative as well as nonrespondent one-subsampling assumes full response after the one-subsampling but two-subsampling schemes do not assume full response until second subsampling phase is completed. In some cases when a nonrespondent one-subsampling scheme is chosen for compensation, some nonrespondent data may not be available. Thus, two-subsampling scheme can be considered for use in compensating nonresponse. Both these two schemes will of course give less variance than when nonrespondents are ignored. Even though in theory one-subsampling schemes are more efficient, in practice two-subsampling schemes are preferable because it is difficult to get full response in the first subsampling phase.

2.2) Weighting Adjustment Procedures: These methods should be used with care. If the response pattern is random, weighting adjustment methods with *RHG* models, namely sample-based, population-based and raking ratio method, can reduce the bias from nonresponse except for unequal probability sampling with

Table 7.1: **Biased** or **Unbiased** Estimator with Compensation Methods in Sampling Designs under **Random Response Mechanism**

Method	Sampling Design									
	srswr	srswor	stwr	stwor	ptwr	ptwor	usrswr	usrswor	ustwr	ustwor
one-sub	1	1	1	1	1	1	1	1	1	1
two-sub	1	1	1	1	1	1	1	1	1	1
naive	2	2	1	1	1	1	2	2	1	1
rhg:g	2	2	-	-	-	-	4	2	-	-
rhg:s	1	1	-	-	-	-	4	1	-	-
rhg:p	1	1	-	-	-	-	4	2	-	-
rhg:r	1	1	-	-	-	-	4	2	-	-
rhg:u	-	-	-	-	-	-	4	-	-	-
sran:n	2	2	1	1	1	1	1	1	1	1
sseq:n	2	2	1	1	1	1	-	-	-	-
sreg:n	2	2	1	1	1	1	1	1	1	1
mran:n	2	2	1	1	1	1	1	1	1	1
mseq:n	2	2	1	1	1	1	-	-	-	-
mreg:n	2	2	1	1	1	1	1	1	1	1
wreg:n	2	2	1	1	1	1	-	-	-	-
abb:n	2	2	1	1	1	1	-	-	-	-
fn:n	2	2	1	1	1	1	-	-	-	-
afn:n	2	2	1	1	1	1	-	-	-	-
mfd:n	2	2	1	1	1	1	-	-	-	-
mwfd:n	2	2	1	1	1	1	-	-	-	-
sran:r	1	1	-	-	-	-	1	1	-	-
sseq:r	1	1	-	-	-	-	-	-	-	-
sreg:r	1	1	-	-	-	-	1	1	-	-
mran:r	1	1	-	-	-	-	1	1	-	-
mseq:r	1	1	-	-	-	-	-	-	-	-
mreg:r	1	1	-	-	-	-	1	1	-	-
wreg:r	1	1	-	-	-	-	-	-	-	-
abb:r	1	1	-	-	-	-	-	-	-	-
fn:r	1	1	-	-	-	-	-	-	-	-
afn:r	1	1	-	-	-	-	-	-	-	-
mfd:r	1	1	-	-	-	-	-	-	-	-
mwfd:r	1	1	-	-	-	-	-	-	-	-

Note: See appendix for notations. Indicators 1, 2, 3 and 4 mean unbiased (0-0.99%), slight biased (1.00-2.99%), more biased (3.00-4.99%) and biased estimator ($\geq$ 5%) respectively. The symbol '-' means the design does not apply for this method, and 'n' and 'r' mean naive and *RHG* model.

Table 7.2: **Biased** or **Unbiased** Estimator with Compensation Methods in Sampling Designs under **Dependent Response Mechanism**

Method	Sampling Design									
	srswr	srswor	stwr	stwor	ptwr	ptwor	usrswr	usrswor	ustwr	ustwor
one-sub	1	1	1	1	1	1	2	1	1	1
two-sub	1	1	1	1	1	1	2	1	1	1
naive	4	4	1	1	1	1	4	4	1	1
rhg:g	4	4	-	-	-	-	4	4	-	-
rhg:s	1	1	-	-	-	-	4	1	-	-
rhg:p	1	1	-	-	-	-	4	1	-	-
rhg:r	1	1	-	-	-	-	4	1	-	-
rhg:u	-	-	-	-	-	-	1	-	-	-
sran:n	4	4	1	1	1	1	2	2	1	1
sseq:n	4	4	1	1	1	1	-	-	-	-
sreg:n	3	3	1	1	1	1	2	2	1	1
mran:n	4	4	1	1	1	1	2	2	1	1
mseq:n	4	4	1	1	1	1	-	-	-	-
mreg:n	2	2	1	1	1	1	2	2	1	1
wreg:n	2	2	1	1	1	1	-	-	-	-
abb:n	4	4	1	1	1	1	-	-	-	-
fn:n	4	4	1	1	1	1	-	-	-	-
afn:n	4	4	1	1	1	1	-	-	-	-
mfd:n	2	2	1	1	1	1	-	-	-	-
mwfd:n	1	1	1	1	1	1	-	-	-	-
sran:r	1	1	-	-	-	-	1	1	-	-
sseq:r	1	1	-	-	-	-	-	-	-	-
sreg:r	1	1	-	-	-	-	1	1	-	-
mran:r	1	1	-	-	-	-	1	1	-	-
mseq:r	1	1	-	-	-	-	-	-	-	-
mreg:r	1	1	-	-	-	-	1	1	-	-
wreg:r	1	1	-	-	-	-	-	-	-	-
abb:r	1	1	-	-	-	-	-	-	-	-
fn:r	1	1	-	-	-	-	-	-	-	-
afn:r	1	1	-	-	-	-	-	-	-	-
mfd:r	1	1	-	-	-	-	-	-	-	-
mwfd:r	1	1	-	-	-	-	-	-	-	-

Note: See appendix for notations.

Table 7.3: **Variance Estimation** with Compensation Methods in Sampling Designs

Method	Sampling Design									
	srswr	srswor	stwr	stwor	ptwr	ptwor	usrswr	usrswor	ustwr	ustwor
one-sub	1	1	1	1	1	1	1	1	1	1
two-sub	1	1	1	1	1	1	1	1	1	1
naive	2	3	2	3	2	3	2	3	2	3
rhg:g	1	1	-	-	-	-	1	1	-	-
rhg:s	1	1	-	-	-	-	1	1	-	-
rhg:p	1	1	-	-	-	-	1	1	-	-
rhg:r	1	1	-	-	-	-	1	1	-	-
rhg:u	-	-	-	-	-	-	1	-	-	-
sran:n	2	2	2	2	2	2	2	2	2	2
sseq:n	2	2	2	2	2	2	-	-	-	-
sreg:n	2	2	2	2	2	2	2	2	2	2
mran:n	2	2	2	2	2	2	2	2	2	2
mseq:n	2	2	2	2	2	2	-	-	-	-
mreg:n	2	2	2	2	2	2	2	2	2	2
wreg:n	2	2	2	2	2	2	-	-	-	-
abb:n	2	2	2	2	2	2	-	-	-	-
fn:n	2	2	2	2	2	2	-	-	-	-
afn:n	2	2	2	2	2	2	-	-	-	-
mfd:n	2	2	2	2	2	2	-	-	-	-
mwfd:n	2	2	2	2	2	2	-	-	-	-
sran:r	1	1	-	-	-	-	1	1	-	-
sseq:r	1	1	-	-	-	-	-	-	-	-
sreg:r	1	1	-	-	-	-	1	1	-	-
mran:r	1	1	-	-	-	-	1	1	-	-
mseq:r	1	1	-	-	-	-	-	-	-	-
mreg:r	1	1	-	-	-	-	1	1	-	-
wreg:r	1	1	-	-	-	-	-	-	-	-
abb:r	1	1	-	-	-	-	-	-	-	-
fn:r	1	1	-	-	-	-	-	-	-	-
afn:r	1	1	-	-	-	-	-	-	-	-
mfd:r	1	1	-	-	-	-	-	-	-	-
mwfd:r	1	1	-	-	-	-	-	-	-	-

Note: See appendix for notations. Indicators 1, 2 and 3 mean reduced variance, increased variance and the same variance respectively.

replacement design. General-based methods do not improve the bias. If the response pattern is dependent on the study variable, the three *RHG* methods, sample-based, population-based and raking ratio methods, are also useful for reducing bias.

The *RHG* methods had lower variance compared with the naive methods. However, if a complex survey design is used, the naive model is adequate. For example, stratified random sampling design with the naive model had lower variance than the *RHG* models in simple random sampling. It is noted that the post-stratified random sampling with the naive model is the population-based adjustment method in simple random sampling (*Oh & Scheuren*, 1983).

2.3 Imputation Methods: In general with **single imputation methods** the stochastic regression imputation can reduce the bias more than random and sequential methods **especially with dependent response**. However, if the three *RHG* models, stochastic regression, random and sequential methods, are used the bias will reduce. For **multiple imputation**, the stochastic regression imputation, and related methods with regression procedures such as modified Wang's regression method and mixture model with the follow-up data or without the follow-up data, can reduce the bias **especially in the dependent response case**. If the **response pattern is random**, all the nine multiple imputation methods can reduce the bias if these methods are used with the *RHG* model.

Comparing the three single imputation methods for large samples, in general, stochastic regression imputation had the lowest variance followed by random imputation and sequential imputation. If the sample is small, stochastic regression imputation can have higher variance than random or sequential imputation. In the case of multiple imputation, in general, modified Wang's regression gave the lowest variance followed by stochastic regression, mixture model without follow-up data, mixture model with follow-up data, adjusted

fully normal, approximated Bayesian bootstrap, fully normal, random and sequential imputation. The mixture model without follow-up data should be used cautiously because this method is sensitive with the CV in the slope coefficient (*Rubin*, 1987).

The unbiasedness and variance reduction for compensation methods compared to ignored nonrespondents are summarised in table 7.1-7.3. For example, if simple random sampling with replacement (*SRSWR*) is used with a naive model and single random imputation, the relative bias is approximately between 1.00%-2.99% for random response mechanism and greater than 5% for dependent response mechanism. The variance of mean estimator with a naive model and single random imputation in *SRSWR* is larger than when nonrespondents are ignored.

7.2 Algorithm for Dealing with Nonresponse

The optimal survey design will depend on the sampling frame, time and budget. The design should include consideration on how to prevent non-sampling errors and in particular nonresponse problems. Methods to reduce nonresponse include call-backs and follow-up techniques. Compensation methods such as nonrespondent subsampling, weighting adjustment and imputation methods should be considered for reducing the bias.

The following algorithm is proposed as a method of assisting in the optimal design and analysis a survey.

1) Survey Plan

   a) If a sampling frame is available and auxiliary information in the frame is categorical, stratified random sampling should be used. If the auxiliary information is quantitative then unequal probability sampling should

be used. If additional categorical information is available then unequal probability sampling should be used with a stratified design. If there is no auxiliary information in the frame, or auxiliary information has low correlation to the study variable then simple random sampling should be used.

b) If simple random sampling were undertaken auxiliary information should, if possible, be collected during the data collection phase. If auxiliary information can be collected post-stratified techniques should be used, otherwise the survey remains a simple random sample.

c) The survey designer should take suitable preventive methods to reduce nonresponse and aim for full nonresponse. For suggested methods such as higher-priority mailing, more call attempts, clearer interviewer assignment materials, special tracing efforts, follow-up reminders, incentives, endorsements, lead letter, proxy respondents, refusal conversion strategies, etc see for example in *Lessler & Kalsbeek* (1992).

2) Once data is collected, if there is nonresponse.

a) If there is no time and cost limitation, nonrespondent subsampling should be used. Again, nonrespondent two-subsampling schemes are preferable (see theorems 3.1-3.12 and tables D.35-D.70).

b) If there is time or cost limitation then compensation methods used during the estimation phase should be used. In general imputation methods are recommended over weighting adjustment methods, because they produce estimates with lower bias. Further, multiple imputation should be used in preference to single imputation if cost and time allow.

If the response mechanism is random then stochastic regression should be used when the sample size is large (see theorems 5.16-5.27 and ta-

bles D.107-D.142, D.161-D.178 and D.197-D.220). If the sample size is small then approximate Bayesian bootstrap is recommended for equal probability sampling (see section 5.3 and tables D.125-D.142) and random imputation for unequal probability sampling (see theorems 5.1-5.12, section 5.3 and tables D.197-D.214). If suitable auxiliary information (a continuous variable, correlated to the study variable) is not available for modelling in stochastic regression and imputation methods are to be used then approximate Bayesian bootstrap, fully normal, adjusted fully normal, random and sequential imputation methods are recommended (see theorems 5.1-5.15, section 5.3 and tables D.107-D.142, D.161-D.178).

If the response mechanism is nonrandom then modified Wang's regression is recommended. Other methods that are also suitable are stochastic regression, mixture model with follow-up data or without follow-up data. (see theorem 5.16-5.27, section 5.3 and tables D.107 -D.262).

c) Weighting adjustment methods should be considered when imputation methods are likely to be very costly. Imputation methods can be very compute intensive for large datasets.

   i) In stratified or simple random sampling with equal probability selection the naive model should be used for without replacement selction (see theorems 4.1-4.4 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104). When equal probability sampling has been used and auxiliary information collected during the survey phase then the sample can be post-stratified, the naive model should be used (see theorems 4.5-4.6 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104). As an alternative rather than post-stratified techniques the *RHG* model can be used. If the *RHG* model is used and there is information on the population size of the cells or classes then the population-based

method should be used (see theorems 4.15-4.16 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104). Note this is equivalent to post-stratified random sampling with the naive model. With the *RHG* model, only information of the cell or class marginal totals then raking ratio should be used (see theorems 4.19-4.20 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104), otherwise sample-based should be used (see theorems 4.11-4.12 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104). However, if there is no information about the post-sample size but known only post-response size, then general-based method should be used. (see theorems 4.23-4.24 and tables D.71-D.76, D.89-D.92, D.95-D.97 and D.104).

ii) In stratified or simple random sampling with unequal probability selection the naive model should be used (see theorems 4.7-4.10 and tables D.77-D.88, D.90-D.91, D.93-D.94, D.98-D.103 and D.105-D.106). If additional categorical information is available from the survey a *RHG* model should be used. For sampling with replacement the bias-removal should be used (see theorem 4.27 and table D.77-D.88, D.90-D.91, D.93-D.94, D.98-D.103 and D.105-D.106) and for sampling without replacement sample-based method should be used (see theorem 4.13 and tables D.77-D.88, D.90-D.91, D.93-D.94, D.98-D.103 and D.105-D.106).

The above steps are illustrated by figure 7.1 and table 7.4. Figure 7.1 shows the process of choosing the survey design. Table 7.4 guides in the choice between weighting adjustment and imputation methods in the analysis phase.

7.3 Summary of Answers to Research Questions

The two research questions were stated as:

1) Which design is best for conducting large-scale surveys such as the Thailand

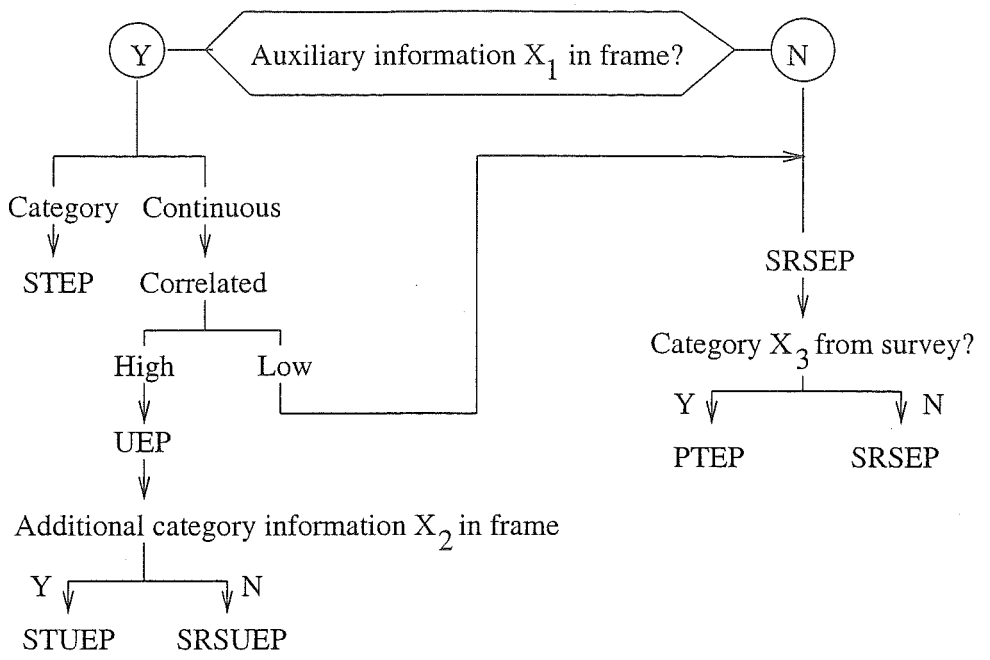


Figure 7.1: Diagram for Deciding on the Survey Design

Note: SRSEP, STEP, PTEP, SRSUEP and STUEP are simple equal probability random sampling, stratified equal probability random sampling, post-stratified equal probability random sampling, simple unequal probability random sampling and stratified unequal probability random sampling respectively.

For the analysis recommended for each of these survey designs see table 7.4

Establishment Survey with nonresponse problems?

2) Which compensation method can reduce the effects from nonresponse problems?

The method and results of the simulation study was summarised in chapter 6. The detailed results are on the accompany compact disk. These results lead to the following answers to the research questions:

1) Sampling Design: If the sampling frame is available with categorised infor-

mation to make sensible strata, stratified random sampling should be used. However, if the frame is not available, post-stratified techniques are recommended. When quantitative auxiliary information in the sampling is highly correlated with the study variable, an unequal probability selection process should be used. If there is no appropriate auxiliary information or not highly correlated auxiliary variable, then an equal probability selection process should be used.

2) Compensation Method: Nonrespondent subsampling which reduces bias and variance is the recommended compensation method. Despite the higher variance two-subsampling scheme is generally preferable to one-subsampling scheme because of the difficulty of getting full response in the first subsampling phase. However, if costs for subsampling are high then compensation in the analysis phase need to be used. Imputation method is preferable to weighting adjustment. For imputation, multiple imputation is recommended rather than single imputation; the variance in single imputation may appear to be lower than in multiple imputation, but this is because the variance in single imputation does not include the variation due to the imputation process. If the sample size is large, one of the regression methods such as modified Wang's regression, stochastic regression, mixture model with or without follow-up data is recommended. If the sample size is small, then one of approximate Bayesian bootstrap, fully normal, adjusted fully normal, random and sequential methods is recommended (see table 7.4). For weighting adjustment, if a suitable auxiliary information can be collected during data collection phase, *RHG* models are recommended, otherwise a naive model should be used. In equal probability sampling, if the strata population sizes are known, then the population-based method should be used. If the data is in a contingency table and marginal population totals are known the raking ratio method is recommended. If the

Table 7.4: Recommended compensation method during the estimation phase varying with survey design

Method	SRSEP	SRSUEP	STEP	STUEP	PTEP
Imputation[3]					
$-Random$	Reg^1	Reg^1	Reg^1	Reg^1	Reg^1
$-Nonrandom$	$Wang^2$	$Wang^2$	$Wang^2$	$Wang^2$	$Wang^2$
Weighting	*Naive* or *RHG* if have X_3 then use *PB* or *RR* if marginal or *SB* if nothing	*Naive* or *RHG* if have X_3 then use *BR* for WR or use *SB* for WOR	*Naive*	*Naive*	*Naive*

Note:

1: If small n then use *approximate Bayesian bootstrap* for equal probability sampling or use *random* imputation for unequal probability sampling.

2: Alternative use is *stochastic regression*, *mixture model with* or *without follow-up data.*

3: If auxiliary information for regression is not available then use *approximate Bayesian bootstrap*, *fully normal*, *adjusted fully normal*, *random* or *sequential.*

Notation: "*Reg*", "*Wang*", "*PB*", "*RR*" and "*SB*" mean *stochastic regression* and *modified Wang's regression*, *population-based*, *raking ratio* and *sample-based* methods respectively.

marginal population totals are not known, then the sample-based method is recommended, otherwise if only response post-stratified sizes are known, then the general-based is recommended. In unequal probability sampling, the recommended weighting adjustment is a bias-removal method for sampling with replacement and a sample-based method for sampling without replacement (see table 7.4).

Appendix A

Thailand

A.1 Geography and Topography

Thailand is in the heart of the Southeast Asian mainland. It covers an area of 513,115 square kilometres. It is bordered by Laos to the northeast, Myanmar to the north and west, Cambodia to the east, and Malaysia to the south. Thailand has maximum dimensions of about 2,500 km north to south and 1,250 km east to west, with a coastline of approximately 1,840 km in the Gulf of Thailand and 865 km along the Indian Ocean.

Thailand is divided into four topographic regions: i) the North, ii) the Central Plain, or Chao Phraya River Basin, iii) the Northeast, or the Korat Plateau, and iv) the South, or Southern Isthmus.

i) The North is a mountainous region characterised by natural forest, ridges and deep, narrow, alluvial valleys.

ii) Central Thailand, the basin of the Chao Phraya River, is a lush, fertile valley. It is the richest and most extensive rice-production area in the country and has

often been called the "Rice Bowl of Asia". Bangkok, the capital of Thailand, is located in this region.

iii) The Northeastern region, or Korat Plateau, is an arid region characterised by a rolling surface and undulating hills. Harsh climatic conditions often result in this region being subjected to floods and droughts.

iv) The southern region is hilly and mountainous, with thick virgin forests and rich deposits of minerals and ores. This region is the centre for the production of rubber and the cultivation of other tropical crops.

A.2 Government

Thailand is governed by a constitutional monarchy with a parliamentarian form of Government. The Bangkok Metropolitan Administration is administrated by an elected Governor and is divided into 38 districts. The country is divided into 76 Provinces, each administered by an appointed Governor, which are sub-divided into districts, sub-districts, tambons (groups of villages) and villages (see figure A.1 and A.2 below).

A.3 Local Administration

In each province, there are three local administration areas: municipal area, sanitary district and out of municipal-sanitary area.

A Municipal area is a legal unit established by the Royal Decree of the 1953 Municipal Act. There are three categories of municipal areas: Nakon (city), Muang (town) and Tambon (Commune)

- A tambon municipal is established whatever it is deemed appropriate.

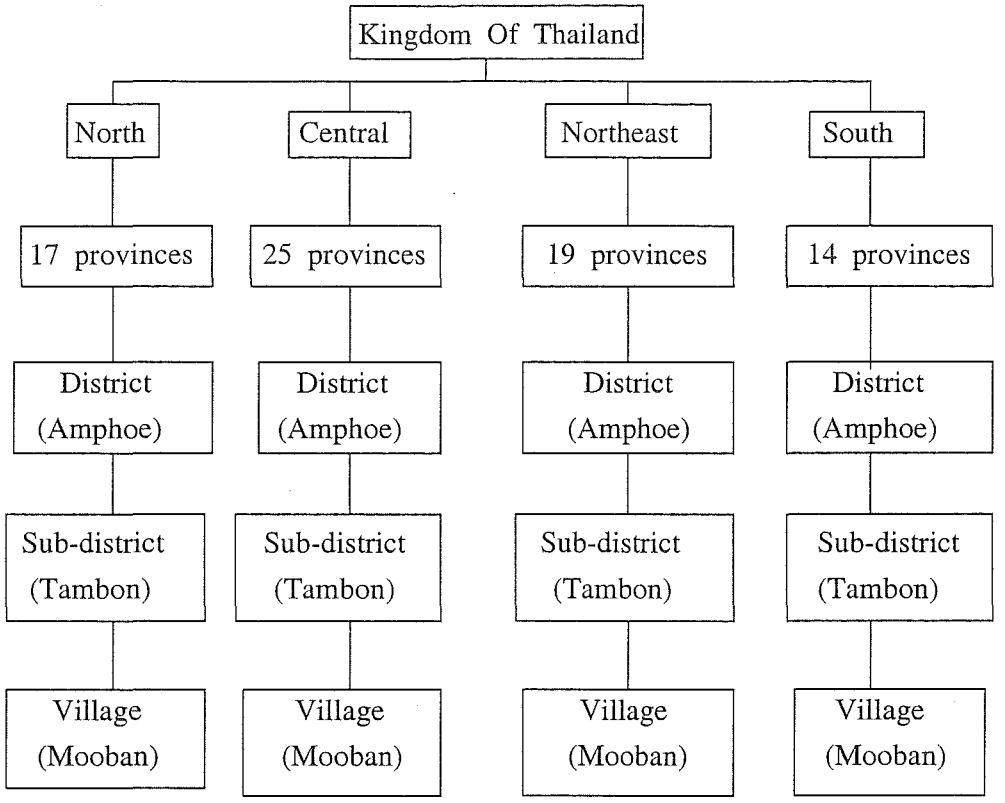


Figure A.1: Structure of Thailand Administration Area

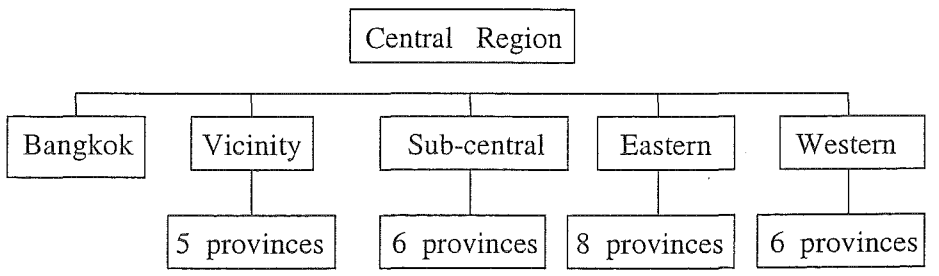


Figure A.2: Structure of The Central Region Administration Area

- A muang municipal is established in each area where the administrative seat of the Provincial Government is located or where the population is at least 10,000 persons, with an average density of not less than 3,000 persons per square kilometre. The sources of tax revenue must also be sufficient for the execution of municipal affairs as stipulated in the 1953 Municipality Act.

- A nakhon municipality is established in area where the population, is at least 50,000 persons, with an average density of not less than 3,000 persons per square kilometre. Tax revenue must also be sufficient for the execution of municipal affairs as stipulated in the 1953 Municipality Act.

A sanitary district is established by the Ministry of Interior under the provisions of the Sanitary District Act of 1952. Under the provisions of the Municipality Act, any sanitary district may be established as a municipal area. There are two categories of sanitary district districts: Urban Sanitary and Rural Sanitary (*National Statistical Office*, 1990).

- Urban Sanitary district refers to the sanitary district where the population is at least 5,000 persons.

- Rural Sanitary district refers to the sanitary district where the population is less than 5,000 persons.

An out of municipal-sanitary district area is the remainder area in the country which is neither municipal area nor sanitary district area.

Appendix B

National Statistical Office

B.1 Organisation of Thailand National Statistical Office

In August 1993, the present organisation of the Thailand National Statistical Office (NSO) was approved by the cabinet. The administration of the NSO is separated into two parts, i.e. the central and provincial administrations. The head of the Office is called the Secretary General. In addition, there are two deputies to the Secretary General, who assist in directing technical and administration works.

The central administration is divided into nine divisions: The Office of the Secretary, Statistical Data Bank and Information Dissemination Division, Field Operation Division, Statistical Policy and Coordination Division, Data Processing Operation Division, Data Processing Techniques Division, Statistical Techniques Division, Economic Statistics Division and Social Statistics Division. The function of the nine division in the central administration, within the headquarter office, are one of the significant components of the statistical system of Thailand. The plans of all the surveys and censuses, including other procedures of statistical works (except for the

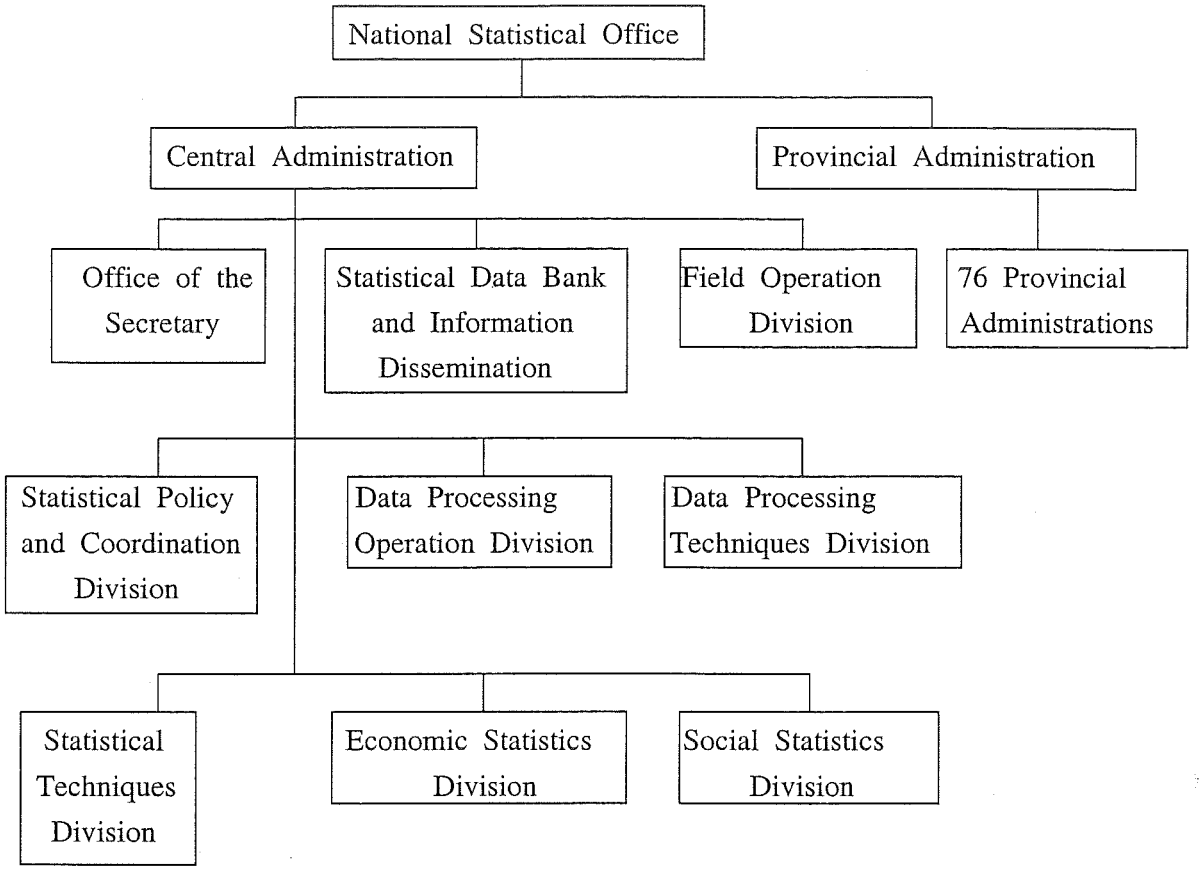


Figure B.1: Structure of Thailand National Statistical Office

field operation), are completed in the headquarter office by any of the nine divisions.

The provincial administration consists of 76 provincial offices, of which their main responsibility is expanded from only the field operation (as what previously performed) to the complete cycle of the statistical works. Furthermore, each provincial office is also expected to perform supporting and supervising roles for the statistical system in its own province.

The organisation chart of the National Statistical Office is summarised in figure B.1

Appendix C

Simulation Program Outline

The basic program structure for the simulations is outline below. There were two advanced steps before a sample was selected. A sampling frame was constructed first with a level of correlation between a study variable Y and an auxiliary variable X condition. Conditions for sampling were then set up such as sample size, selection procedure and response pattern.

C.1 Program outline for simulations

There were nine steps for simulation in this study:

i) Construct a sampling frame.

ii) Set up sample conditions.

iii) Choose a sample survey design.

iv) Select sample units from a sampling frame.

v) Indicate response/nonresponse unit in a sample.

vi) Compute a mean and a variance for sample units and response units.

vii) Choose a compensation method.

viii) Compute a mean and a variance when using a compensation method in step 7.

ix) Repeat step iv-viii 1,000 times and then average the mean and the variance computed from step vi and viii.

Notes: Details for i)-v) and vii) are presented separately in section C.2-C.7 respectively.

C.2 Sampling Frame

A sampling frame consisted of a study variable Y, auxiliary variables such as a quantitative variable X using for unequal probability sampling , and two qualitative variable Z_1 and Z_2 using for stratified random sampling and post-stratified random sampling. It also included many indicators: W for post-stratified process, I for a sample process and the others which related to sampling unit links for sample designs and selection processes.

C.3 Sample Conditions

There were three conditions used to select a sample from a sampling frame: i) sample size, ii) selection procedure and iii) response pattern.

i) Three sample size were used for this simulation study: 15%, 30% and 50% of population size.

ii) Two selection procedures were used for this simulation study: sampling with replacement and sampling without replacement.

iii) Two response patterns were used for this simulation study: random response and dependent response with the study variable Y. There were five levels of random response: low response (10%), medium response (30% , 50% and 70%) and high response (90%).

C.4 Sample Survey Design

There were ten basic one-stage survey design based on equal/unequal probability sampling, and sampling with/without replacement: *SRSWOR*, *SRSWR*, *STWOR*, *STWR*, *PTWOR*, *PTWR*, *USRSWOR*, *USRSWR*, *USTWOR* and *USTWR*.

C.5 Select Sample Units

The aim in indicating a chosen sampling unit was to distinguish between units in a population to show whether the unit was selected or not. This step was helpful for sampling without replacement scheme. There were two algorithms depending on the survey design: i) equal probability selection procedure and ii) unequal probability selection procedure.

i) Equal probability selection algorithm:

1) Retrieve sample indicator I, whose value is zero, from a sampling frame.

2) Select random integer number between 1 and N, say k.

3) Check sample indicator, I_k, for sampling with/without replacement

3.1) If draw a sample first time, skip this check and go to step 4.

3.2) Check I_k by using case 1 or 2:

case 1: sampling without replacement

If $I_k = 0$, select this unit; otherwise back to step 2.

case 2: sampling with replacement

Skip this check.

4) Replace value 1 into sample indicator I_k.

5) Repeat step 2-4 until sample size equals n.

ii) Unequal probability selection algorithm:

1) Retrieve sample indicator I, whose value is zero, from a sampling frame.

2) Create two variables to use for indicating a sampling unit chosen: Amin and Amax. With X is an integer quantitative variable,

2.1) $Amin(j) = Amin(j-1) + X(j-1) + 1$ for $j = 2, ..., N$; where $Amin(0) = 0$.

2.2) $Amax(j) = Amax(j-1) + X(j)$ for $j = 1, ..., N$; where $Amax(0) = 0$.

3) Select random integer number between 1 and Amax(N), then identify sampling unit k.

4) Check sample indicator, I_k, for sampling with/without replacement.

4.1) If draw a sample first time, skip this check and go to step 5.

4.2) Check I_k by using case 1 or 2:

case 1: sampling without replacement

If $I_k = 0$, select this unit; otherwise go back to step 3.

case 2: sampling with replacement

Skip this check.

5) Replace value 1 into sample indicator I_k.

6) Repeat step 3-5 until sample size equals n.

C.6 Response/Nonresponse Units

There were two algorithms to indicate for a unit in the sample whether the response was random or dependent on Y.

i) Random response algorithm:

1) Create a response indicator R, whose value is one, in a sample.

2) Generate a random variable U whose distribution is uniform (0,1).

3) Compare a value of U in sample unit k with a response rate:

If $U_k \leq$ a fixed response rate, then indicate response sample $R_k = 1$ for k=*1,...,n*; otherwise $R_k = 0$ indicating nonresponse in unit k.

4) Repeat step 2-3 until $k = n$.

ii) Dependent response with Y algorithm:

1) Create a response indicator R, whose value is one, in a sample.

2) Set up two values of study variable: Y_a and Y_b where $Y_a < Y_b$.

3) Compare Y_k in a sample unit k for k=*1,...,n* by using 3.1. 3.2 or 3.3:

3.1) If $Y_k \leq Y_a$, then indicate unit k is respond, $I_k = 1$.

3.2) If $Y_k > Y_b$, then

3.2.1) Generate a random variable U whose distribution is uniform (0,1).

3.2.2) Compare a random variable U with a fixed response rate: If $U_k \leq$ a fixed response rate, then indicate unit k is respond ($I_k = 1$); otherwise $I_k = 0$ indicating nonresponse in unit k.

3.3) if $Y_a < Y_k \leq Y_b$, then

3.3.1) Generate a random variable U whose distribution is uniform (0,1).

3.3.2) Compare a random variable U with a response function $f(Y_k) = a + bY_k$ as shown in figure C.1: If $U_k \geq f(Y_k)$, then indicate unit k is respond ($I_k = 1$); otherwise $I_k = 0$ indicating non-response in unit k.

4) Repeat step 2-3 until $k = n$.

C.7 Compensation Methods

There were three general methods used for compensating nonresponse data: nonrespondent sabsampling, weighting adjustment procedure, and imputation methods.

C.7.1 Nonrespondent subsampling algorithm

There were two nonrespondent subsampling schemes: i) one-subsampling and ii) two-subsampling:

i) one-subsampling scheme procedure

1) Check a nonrespondent sample unit k for k=*1,...,n*

   If $R_k = 0$, then generate a random variable V whose distribution is uniform (0,1); otherwise skip to step 3.

2) If $V < 0.5$, then a sample unit k is selected for nonrespondent subsampling.

3) Repeat step 1-2 until nonrespondent subsample size $= n'_{12}$, where $n'_{12} = \frac{n_{12}}{k}$ and k is an integer which is greater than 0, then return to the main program.

ii) two-subsampling scheme procedure

1) Create a subsampling indicator $S_k = 0$ for k=*1,...,n*.

2) Check a nonrespondent sampling unit k for k=*1,...,n*: If $R_k = 0$, then generate a random variable V whose distribution is uniform (0,1); otherwise skip to step 5.

3) If $V < 0.5$, then unit k is selected for first subsampling and replace 1 into S_k; otherwise replace 2 into S_k.

4) Check a number of first subsampling: If first subsampling count $\geq n'_{12}$, then return to the main program (in this case two-subsampling scheme is reduced to one-subsampling scheme).

5) Repeat step 2-4 until first subsampling size $= n_{21}$.

6) Check a second subsampling for k=*1,...,n*: If $S_k = 2$, then generate a random variable U whose distribution is uniform (0,1); otherwise skip to step 8.

7) If $U < 0.5$, then a nonresponse unit k is selected for second subsampling.

8) Repeat step 6-7 until second subsampling size $= n'_{22}$, then return to the main program.

C.7.2 Weighting adjustment algorithm

Check a sample design to choose a response model as:

1) If a sample design is simple random sampling, then choose a response model between naive and *RHG*; otherwise use a naive model (for stratified random sampling and post-stratified random sampling)

2) Set up a weighting factor to response units depend on method used in equal/unequal probability sampling with/without replacement (more details and formulae used in chapter 4).

Note: For a *RHG* model in simple random sampling, post-stratification technique is used to identify a sample unit into an exact post-stratum.

C.7.3 Imputation method algorithm

There were two general imputation methods used for missing data: i) single imputation and ii) multiple imputation:

i) Single imputation procedure

1) Choose an imputation method to replace imputed value into nonresponse units.

2) Compute imputed value unit k for k=*1,...,r*.

3) Store value unit k and repeat step 2 until k=r, then return to main program.

ii) Multiple imputation procedure

1) Choose an imputation method to replace imputed value into nonresponse units.

2) Compute imputed value unit k for $k=1,\ldots,r$.

3) Store value unit k and repeat step 2 until k=r.

4) Repeat step 2-3 M times, then return to main program.

Nine imputation methods were used to impute value; (1) random, (2) sequential, (3) stochastic regression, (4) Wang's regression, (5) approximate Bayesian bootstrap, (6) fully normal, (7) adjusted normal, (8) mixture model with follow-up data and (9) mixture model without follow-up data. The first three methods were used to apply both single and multiple imputation. Methods 4-10 were used only in multiple imputation. Algorithm for random, sequential and stochastic regression are described here. However, algorithms in method 4-9 are presented in chapter 5.

i) Random imputation procedure

1) Store m response units and set up indicator, $B_k = 1$ for $k=1,\ldots,m$ for these units.

2) Compare between size of response units (m) and of nonresponse units (r): If $m < r$, random imputation without replacement case cannot be applied (case random response: 10%, 30% and 50%).

3) Generate integer number between 1 and m, say k.

4) Check sampling with/without replacement for response unit k

4.1) If survey design is sampling with replacement, replace 0 to B_k and skip to step 5; otherwise check

4.2) If $B_k = 0$ skip to step 6; otherwise go to step 5.

5) Replace value of random variable Y of unit k to imputed value i for $i=1,\ldots,r$ and replace 0 to B_k.

6) Repeat step 3-5 until $i=r$.

ii) Sequential imputation procedure

1) Shuffle a sample of size n which includes response units and nonresponse units.

2) Check the first sample unit ($k=1$)

If $R_1 = 1$, then store value Y_1 to donor register D and skip to step 3; otherwise average value Y of response units and replace the average to the sample unit 1 and store it to D.

3) Check response units for $k=2,\ldots,n$: If $R_k = 1$ then store value Y_k to D and go to step 4; otherwise replace value D to a sample unit k.

4) Repeat step 3 until $k=n$.

iii) Stochastic imputation procedure

1) Calculate ordinary least square regression coefficients $\hat{\beta}_{0m}$ and $\hat{\beta}_{1m}$ from response data.

2) Retrieve X from a sampling frame for nonrespondents.

3) Randomly select the $r = n - m$ observed residuals from $\{e_k = y_k - \hat{\beta}_{0m} - \beta_{1m} x_k\}$ for k=*1,...,r*.

4) Compute the stochastic regression imputation by using $\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k + e_k$ for k=*1,...,r*.

5) Repeat step 3-4 until k=*r*.

Appendix D

List of Tables in Compact Disk

This appendix outlines the tables which summarise the simulation results. The results are the relative bias, the coefficient of variation of mean estimator and the design effect. There are six Excel worksheet files in the CD named cd1, cd2, cd3, cd4, cd5 and cd6.

The cd1 consists of 3 worksheets as: i) Sheet 1 consists of Table D.1-D.34, ii) Sheet 2 consists of Table D.35-D.58 and iii) Sheet 3 consists of Table D.59-D.70.

The cd2 consists of 2 worksheets as: i) Sheet 1 consists of Table D.71-D.94 and ii) Sheet 2 consists of Table D.95-D.106.

The cd3 consists of 2 worksheets as: i) Sheet 1 consists of Table D.107-D.124 and ii) Sheet 2 consists of Table D.125-D.160.

The cd4 consists of 4 worksheets as: i) Sheet 1 consists of Table D.161-D.178, ii) Sheet 2 consists of Table D.179-D.196, iii) Sheet 3 consists of Table D.197-D.214 and iv) Sheet 4 consists of Table D.215-D.220.

The cd5 consists of 4 worksheets as: i) Sheet 1 consists of Table D.221-D.232, ii) Sheet 2 consists of Table D.233-D.238, iii) Sheet 3 consists of Table D.239-D.250 and iv) Sheet 4 consists of Table D.251-D.262.

The cd6 consists of 3 worksheets as: i) Sheet 1 consists of Table E.1-E.8, ii) Sheet 2 consists of Table E.9-E.16 and iii) Sheet 3 consists of Table E.17-E.24.

Appendix E

Notation and Symbol

The following notations and symbols are used:

abb: Approximated Bayesian bootstrap

abb:n: Multiple approximate Bayesian bootstrap imputation with naive model

abb:r: Multiple approximate Bayesian bootstrap imputation with *RHG* model

afn: Adjusted fully normal method

afn:n: Multiple adjusted fully normal imputation with naive model

afn:r: Multiple adjusted fully normal imputation with *RHG* model

CV: Coefficient of variation

d: Design

deff: Design effect

depwr: Sampling with replacement in dependent response mechanism

depwor: Sampling without replacement in dependent response mechanism

EPSEM: Equal probability selection method

fn: Fully normal method

fn:n: Multiple fully normal imputation with naive model

fn:r: Multiple fully normal imputation with *RHG* model

g: General-based adjustment method

high: High level of correlation between a study variable and auxiliary information

husrs: High level of correlation in unequal probability simple random sampling

hust: High level of correlation in unequal probability stratified random sampling

IID: Independent identically distributed

k_i: a predetermined value for i^{th} nonrespondent subsampling

low: Low level of correlation between a study variable and auxiliary information

lusrs: Low level of correlation in unequal probability simple random sampling

lust: Low level of correlation in unequal probability stratified random sampling

m: Method

mar: Missing at random

mcar: Missing completely at random

meff: Misspecification effect

mfd:n: Multiple mixture model with follow-up data with naive model

mfd:r: Multiple mixture model with follow-up data with *RHG* model

mran:n: Multiple random imputation with naive model

mran:r: Multiple random imputation with *RHG* model

mreg:n: Multiple stochastic regression imputation with naive model

mreg:r: Multiple stochastic regression imputation with *RHG* model

mse: Mean square error

mseq:n: Multiple sequential imputation with naive model

mseq:r: Multiple sequential imputation with *RHG* model

mwfd:n: Multiple mixture model without follow-up data with naive model

mwfd:r: Multiple mixture model without follow-up data with *RHG* model

n: Initial random sample size

n_{ij}: Sample size for inital sample ($i = 1$), first subsampling ($i = 2$), response group ($j = 1$) and nonresponse group ($j = 2$)

$n_{ij}^{'}$: Size of i^{th} nonresponse subsampling

na: Naive model

nse: Nonsampling error

NSO: National Statistical Office

one-sub: Nonrespondent one-subsampling scheme

opt: Optimum

p: Population-based adjustment method

pps: Unequal probability sampling with replacement

pt: Post-stratified random sampling

ptwr: Post-stratified random sampling with replacement

ptwor: Post-stratified random sampling without replacement

pt-one: Nonrepondent one-subsampling scheme in post-stratifed random sampling

pt-two: Nonrepondent two-subsampling scheme in post-stratified random sampling

PES: Post Enumeration Survey

PPS: Probability proportional to size

r: Raking ratio adjustment method

rhg:g: General-based adjustment method

rhg:s: Sample-based adjustment method

rhg:p: Population-based adjustment method

rhg:r: Raking ratio adjustment method

rhg:u: bias-removal adjustment method

R: Response set

RHG: Random homogeneity group model

R_{hl}: Response set in stratum hl

s: Sample-based adjustment method

s_{i2}^2: Sample response variance for i^{th} nonrespondent subsampling

s_m^2: Sample total response variance

se: Sampling error

sel: Selection method

sran:n: Single random imputation with naive model

sran:r: Single random imputation with *RHG* model

sreg:n: Single stochastic regression imputation with naive model

sreg:r: Single stochastic regression imputation with *RHG* model

srs: Simple random sampling

srswr: Simple random sampling with replacement

srswor: Simple random sampling without replacement

srs-one: Nonrepondent one-subsampling scheme in simple random sampling

srs-two: Nonrepondent two-subsampling scheme in simple random sampling

sseq:n: Single sequential imputation with naive model

sseq:r: Single sequential imputation with RHG model

st: Stratified random sampling

stwr: Stratified random sampling with replacement

stwor: Stratified random sampling without replacement

st-one: Nonrepondent one-subsampling scheme in stratified random sampling

st-two: Nonrepondent two-subsampling scheme in stratified random sampling

S: Sample set

S^2 or σ^2: Population variance

S_{hl}: Sample set in stratum hl

two-sub: Nonrespondent two-subsampling scheme

u: bias-removal adjustment method

usrswr: Unequal probability selection on simple random sampling with replacement

usrswor: Unequal probability selection on simple random sampling without replacement

ustwr: Unequal probability selection on stratified random sampling with replacement

ustwor: Unequal probability selection on stratified random sampling without replacement

var: Variance

wor: Without replacement

wr: With replacement

wreg:n: Multiple Wang's modified regression imputation with naive model

wreg:r: Multiple Wang's modified regression imputation with *RHG* model

y: Sum of random variable Y

y_{ij}: Sum of random variable Y for inital sample ($i = 1$), first subsampling ($i = 2$), response group ($j = 1$) and nonresponse group ($j = 2$)

y'_{ij}: Sum of random variable Y of i^{th} nonresponse subsampling and response group ($j = 1$) or nonresponse group ($j = 2$)

πps: Unequal probability sampling without replacement

#ptwr: Post-stratified random sampling with replacement with #% sample size

#ptwor: Post-stratified random sampling without replacement with #% sample size

#srswr: Simple random sampling with replacement with #% sample size

#srswor: Simple random sampling without replacement with #% sample size

#stwr: Stratified random sampling with replacement with #% sample size

#stwor: Stratified random sampling without replacement with #% sample size

#wr: Sampling with replacement with #% sample size

#wor: Sampling with replacement without #% sample size

#-wr-full: Sampling with replacement with #% sample size and full response

#-wr-ignored: Sampling with replacement with #% sample size and ignored nonrespondents

#-wr-one: Nonrespondent one-subsampling scheme with #% sample size in sampling with replacement

#-wr-two: Nonrespondent two-subsampling scheme with #% sample size in sampling with replacement

#-wor-full: Sampling without replacement with #% sample size and full response

#-wor-ignored: Sampling without replacement with #% sample size and ignored nonrespondents

#-wor-one: Nonrespondent one-subsampling scheme with #% sample size in sampling without replacement

#-wor-two: Nonrespondent two-subsampling scheme with #% sample size in sampling without replacement

$\hat{\mu}_m^d$: Sample mean of random variable Y in sampling design "d" and model "m"

$\hat{\mu}_{ij}$: Estimated mean of random random variable Y for initial sample ($i = 1$), first subsampling ($i = 2$) with response group ($j = 1$) or nonresponse group ($j = 2$)

$\hat{\mu}_{ij}^{'}$ Estimated mean of random variable Y for i^{th} nonrespondent subsampling with response group ($j = 1$) or nonresponse group ($j = 2$)

$\hat{\tau}_m^{d,sel}$: Sample total of random variable Y in sampling design "d", model "m" and selection method "sel"

Bibliography

[1] Armitage, P.A. and Colton, T. (1998) *The Encyclopedia of Biostatistics*, John Wiley: Chichester.

[2] Bailar, B.A., Bailey, L. and Corby, C. (1978), Comparison of Some Adjustment and Weighting Procedures for Survey Data, pp 175-198 In: *Survey Sampling and measurement*, ed, Namboodiri N.K., Academic Press:New York.

[3] Barnard, J., Rubin, D.B. and Schenker, N. (1998), Multiple Imputation: In *The Encyclopedia* Eds, Armitage, P.A. and Colton, T., pp 2773-2780.

[4] Basu, D. (1971), An Essay on the Logical Foudations of Survey Sampling, Part I, pp 203-239. In:*Foundation of Statistical Inference: A symposium*, Eds: Godambe, V.P. and Sprott D.A., Holt, Rinehart and Winster of Canada: Toronto.

[5] Bethlehem, J.G. (1988), Reduction of nonresponse bias through regression estimation, *Journal of Official Statistics*, 4, 251-260.

[6] Biemer, P.P. (1991) *Measurement Errors in Surveys*, John Wiley:New York.

[7] Brackstone, G.J. and Rao, J.N.K. (1976), Raking Ratio Estimator, *Survey Methodology*, 2, 63-69.

[8] Brewer, K.R.W. and Hanif, M. (1983) *Sampling with Unequal Probabilities*, Springer-Verlag:New York.

[9] Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977) *Foundation of Inference in Survey Sampling*, John Wiley:New York.

[10] Chaudhuri, A. and Stenger, H. (1992) *Survey Sampling*, Marcel-Dekker:New York.

[11] Chaudhuri, A. and Vos, J.W.E. (1988) *Unified Theory and strategies of Survey Sampling*, North-Holland Publishing:New York.

[12] Cochran, W.G. (1963) *Sampling Techniques. 2rd edition*, John Wiley:New York.

[13] Cochran, W.G. (1977) *Sampling Techniques. 3rd edition*, John Wiley:New York.

[14] Colledge, M.J., Johnson, J.H., Pare, R.M. and Sande, I.G. (1978), Large Scale Imputation of Survey Data, *Survey Methodology*, 4, 203-223.

[15] Cornfield, J. (1951), Modern methods in the Sampling of Human Populations, *American Journal of Public Health*, 41:654-661.

[16] Dagpunar, J. (1988), *Principles of Random Variate Generation*, Clarendon Press:Oxford.

[17] Dalenius, T. (1974), Ends and Means of Total Survey Design. *Forskningsprojektet Fel i Undersokningar*, Stockholm: University of Stockholm.

[18] David, M.H., Little, R.J.A., Samuhel, M.E., and Triest, R.K. (1986), Alternative Methods for CPS Income Imputation, *Journal of the American Statistical Association*, 81, 29-41.

[19] Deming, W.E. (1950), *Some Theory of Sampling*, Dover:New York.

[20] Deming, W.E. (1953), On a Probability Mechanism to Attain an Economic balance between the Resultant Error of Nonresponse and the Bias of Nonresponse, *Journal of the American Statistical Association*, 48, 743-772.

[21] Deming, W.E. and Stephan, F.F. (1940), On a Least Squares adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known, *Annals of Mathematical Statistics*, 11, 427-444.

[22] Deville, J.C., Sarndal, C.E. and Sautory, O. (1993), Generalized Raking Procedures in Survey Sampling, *Journal of the American Statistical Association*, 88, 1013-1020.

[23] Dillman, D.A. (1998), Call-backs and Mail-backs in Sample Surveys: In *The Encyclopedia* Eds, Armitage, P.A. and Colton, T., pp 466-468.

[24] El-Badry, M.E. (1956), A Sampling Procedure for Mailed Questionnaires, *Journal of the American Statistical Association*, 51, 209-227.

[25] Ericson, W.A. (1967), Optimal Sample Design With Nonresponse, *Journal of the American Statistical Association*, 62, 63-78.

[26] Fay, R.E. (1996), Alternative Paradigms for the Analysis of Imputed Survey Data, *Journal of the American Statistical Association*, 91, 490-498.

[27] Ford, B.L. (1983), An Overview of Hot-Deck Procedures, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[28] Foreman, E.K. (1991) *Survey Sampling Principles*, Marcel-Dekker:New York.

[29] Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (1995), *Statistical Inference*, Prentice Hall:New York.

[30] Glynn, R.J., Laird, N.M. and Rubin, D.B. (1986), Mixture Modelling Versus Selection Modelling with Nonignorable Nonresponse, In *Drawing Inferences from Self-selected Samples*, Eds: Wainer, H., Springer-Verlag: New York

[31] Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993), Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups, *Journal of the American Statistical Association*, 88, 984-993.

[32] Godambe, V.P. and Sprott, D.A. (1971) *Foundation of Statistical Inference: A Symposium*, Holt, Rinehart and Winster of Canada:Toronto.

[33] Godambe, V.P. and Thompson, M.E. (1988), One Single Stage Unequal Probability Sampling, pp 111-122 In: *Handbook of Statistics 6: Sampling*, Eds: Krishanaiah, P.R. and Rao, C.R., North-holland Publishing:New York.

[34] Govindarajulu, Z. (1999) *Elements of Sampling Theory and Methods*, Prentice Hall:New York.

[35] Greenless, J.S., Reece, W.S. and Zieschang, K.Y. (1982), Imputation with missing values, *Journal of the American Statistical Association*, 50, 417-432.

[36] Groves, R.M. (1984) *Survey Errors and Survey Costs*, John Wiley:New York.

[37] Hansen, M.H. and Hurwitz, W.N. (1946), The Problem of Non-Response in Surveys, *Journal of the American Statistical Association*, 41, 517-529.

[38] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory* Vols. I and II, John Wiley:New York.

[39] Haslett, S.J. (1985), The Linear Non-Homogeneous Estimator in Sample Survey, *Sankhya*, Series B, 47, 101-117.

[40] Heitjan, D.F. and Rubin, D.B. (1990), Inference From Coarse Data Via Multiple Imputation With Application to Age Heaping, *Journal of the American Statistical Association*, 85, 304-314.

[41] Herzog, T.N., Rubin, D.B. (1983), Using Multiple Imputations to Handle Nonresponse in Sample Surveys, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[42] Hidiroglou, M.A. and Drew, J.D. (1993), A Frame for Measuring and Reducing Nonresponse in Surveys, *Survey Methodology*, 19, 81-94.

[43] Holt, D. and Elliot, D. (1991), Methods of Weighting for Unit Non-Response, *The Statistician*, 40, 333-342.

[44] Holt, D. and Smith, T.M.F. (1979), Post-Stratification, *Journal of the Royal Statistical Society: Series A*, 142, 33-45.

[45] Horvitz, D.G. and Thompson, D.J. (1952), A Generaization of Sampling Without Replacement from a Finite Population, *Journal of The American Statistical Association*, 47, 663-685.

[46] Ireland, C.T. and Kullback, S. (1968), Contingency Tables With Given Marginals, *Biometrika*, 55, 179-188.

[47] Jagers, P. (1986), Post-Stratification against Bias in Sampling, *International Statistical Review*, 54, 159-167.

[48] Jagers, P., Oden, A. and Trulsson, L. (1985), Post-Stratification and Ratio Estimation: Usages of Auxiliary Information in Survey Sampling and Opinion Polls, *International Statistical Review*, 53, 221-238.

[49] Jessen, R.J. (1978) *Statistical Survey Techniques*, John Wiley:New York.

[50] Jinn, J.H., Albany, S. and Sedransk, J. (1989a), Effect on Secondary Data Analysis of the USe of Imputed Values: The Case Where Missing Data Are Not Missing At Random, In: Proceedings of The Survey Research Methods Section, *American Statistical Association*, 509-530.

[51] Jinn, J.H. and Sedransk, J. (1989b), Effect on Secondary Data Analysis of the Use of Imputed Values: The Case Where Missing Data Are Not Missing At Random, *Proceedings of the Survey Research Methods Section, American Statisitcal Association*, 51-61.

[52] Kalton, G. (1983), Models in the Practice of Survey Sampling, *International Statistical Review*, 51, 175-188.

[53] Kalton, G., Kasprzyk, D. (1982), Imputing for Missing Survey Responses, *Proceedings of the Section for Survey Research Methods, American Statistical Association*, 22-33.

[54] Kalton, G., Kasprzyk, D. (1986), The Treatment of Missing Data. *Survey Methodlogy*, 12, 1-16.

[55] Kish, L. (1965), *Survey Sampling*, John Wiley:New York.

[56] Kish, L. (1992), Weighting for Unequal P_i, *Journal of Official Statistics*, 8, 183-200.

[57] Kish, L. (1995), Methods for Design Effects, *Journal of Official Statistics*, 11, 55-77.

[58] Konijn, H.S. (1986) *Statistical Theory of Sample Survey Design and Analysis*, North-holland Publishing:New York.

[59] Krishanaiah, P.R. and Rao, C.R. (1988) *Handbook of Statistics 6: Sampling*, North-holland Publishing:New York.

[60] Kviz, F. (1998), Nonresponse in Sample Surveys: In *The Encyclopedia* Eds, Armitage, P.A. and Colton, T., pp 3043-3047.

[61] Lehtonen, R. and Pahkinen, E.J. (1995) *Practical Methods for Design and Analysis of Complex Survey*, John Wiley: New York.

[62] Lessler, J.T. and Kalsbeek, W.D. (1992) *Nonsampling Error in Surveys*, John Wiley: New York.

[63] Levy, P.S. and Lemeshow, S. (1999) *Sampling of Populations: Methods and Applications. 3rd edition*, John Wiley:New York.

[64] Li, K.H., Raghunathan, T.E. and Rubin, D.B. (1991), Large-Sample Significance levels From Multiply Imputed Data Using Moment-Based Statistics and F Reference Distribution, *Journal of the American Statistical Association*, 86, 1065-1073.

[65] Little, R.J.A. (1982), Models for Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, 77, 237-250.

[66] Little, R.J.A. (1986), Survey Nonresponse Adjustments for Estimates of Means, *International Statistical Review*, 54, 139-157.

[67] Little, R.J.A. (1988), Missing Data Adjustment in large Surveys. *Journal of Business and Economic Statistics*, 6, 287-301.

[68] Little, R.J.A. (1993), Post-Stratification: A Modeler's Perspective, *Journal of the American Statistical Association*, 88, 1001-1012.

[69] Little, R.J.A. (1998), Biostatistical Analysis With Missing Data: In *The Encyclopedia* Eds, Armitage, P.A. and Colton, T., pp 2622-2635.

[70] Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*, John Wiley:New York.

[71] Lohr S.H. (1999) *Sampling: Design and Analysis*, Duxbury Press:New York.

[72] Madow, W.G. and Olkin, I. (1983) *Incomplete Data in Sample Survey Volume1: Report and Case Studies*, Academic Press:New York.

[73] Madow, W.G. and Olkin, I. (1983) *Incomplete Data in Sample Survey Volume2: Theory and Bibliographies*, Academic Press:New York.

[74] Madow, W.G. and Olkin, I. (1983) *Incomplete Data in Sample Survey Volume3: Proceedings of the Symposium*, Academic Press:New York.

[75] Meng, X.L. and Rubin, D.B. (1992), Performing likelihood Ratio Tests With Multiply Imputed Data Sets, *Biometrika*, 79, 811-822.

[76] Mood A.M., Graybill, F.M. and Boes D.C. (1974) *Introduction to the Theory of Statistics 3rd edition*, MaGraw-Hill:New York.

[77] Murthy, M.N. (1967) *Sampling Theory and Methods*, Statistical Publishing Society:New Deli.

[78] Namboodiri, N.K. (1978) *Survey Sampling and measurement*, Academic Press:New York.

[79] Nordholt, E.S. (1998), Imputation: Methods, Simulation Experiments and Practical Examples, *International Statistical Reviews*, 66, 157-180.

[80] National Statistical Office of Thailand (1990), *Report of the 1990 Census of Population and Housing*, Office of the Prime Minister:Bangkok.

[81] National Statistical Office of Thailand (1995), *Report of the 1995 Industrial Survey*, Ofice of the Prime minister:Bangkok.

[82] National Statistical Office of Thailand (1996), *Report of the 1996 Household Labor Force Survey*, Ofice of the Prime minister:Bangkok.

[83] Oh, H.L. and Scheuren, F.J. (1983), Weighting Adjustment for Unit Nonresponse, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[84] Palit, C.D. and Guttman, I. (1973), Bayesian Estimation Procedures for Finite Populations, Single Stage Design, and Normal Populations, *Communications in Statistics*, 1, 93-111.

[85] Politz, A., Simmons, W. (1949), An Attempt to get the "Not-at-Home" into the Sample Without Call-backs, *Journal of the American Statistical Association*, 44, 9-31.

[86] Raj, D. (1956) Some Estimates in Sampling with varying Probabilities Without Replacement, *Journal of the American Statistical Association*, 61, 391-397.

[87] Raj, D. (1968) *Sampling Theory*, McGraw-Hill: New York.

[88] Rao, J.N.K (1968), Some Nonresponse Sampling Theory When the Frame Contains an Unknown Amount of Duplication, *Journal of the American Statistical Association*, 63, 87-90.

[89] Rao, J.N.K. (1973), On Double Sampling For Stratification and Analytical Surveys, *Biometrika*, 60, 125-133.

[90] Rao, J.N.K. (1985), Conditional Inference in Survey Sampling, *Survey Methodology*, 11, 15-31.

[91] Rao, J.N.K. (1996), On Variance Estimation With Imputed Survey Data, *Journal of the American Statistical Association*, 91, 499-520.

[92] Rao, J.N.K. and Bellhouse, D.R. (1990), History and Development of the Theoretical Foundations of Survey Based Estimation and Analysis, *Survey Methodology*, 16, 3-29.

[93] Rao, J.N.K. and Ghangurde, P.D. (1972), Bayesian Optimization in Sampling Finite Populations, *Journal of the American Statistical Association*, 67, 439-443.

[94] Rao, J.N.K. and Hughes, E. (1983) Comparison of Domains in the Presence of Nonresponse, In: *Incomplete Data in Sample Survey, Volume 3*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[95] Rao, P.S.R.S. (1983), Randomization Approach, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[96] Rossi, P.H., Wright, J.D. and Anderson, A.B. (1983) *Handbook of Survey Research*, Academic Press:New York.

[97] Rubin, D.B. (1976), Inference and Missing Data, *Biometrika*, 63, 581-592.

[98] Rubin, D.B. (1977), Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys, *Journal of the American Statistical Association*, 72, 538-543.

[99] Rubin, D.B. (1981), The Bayesian Bootstrap, *The Annals of Statistics*, 9, 130-134.

[100] Rubin, D.B. (1983), Conceptual Issues in the Presence of Nonresponse, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[101] Rubin, D.B. (1986), Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations, *Journal of Business and Economic Statistics*, 4, 87-94.

[102] Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Survey*, John Wiley: New York.

[103] Rubin, D.B. (1996), Multiple Imputation After 18+ Years, *Journal of the American Statistical Association*, 91, 473-489.

[104] Rubin, D.B. and Schenker, N. (1986), Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse, *Journal of the American Statistical Association*, 81, 366-374.

[105] Rubin, D.B. and Schenker, N. (1987), Interval Estimation from Multiple Imputed Data: A Case Study Using Agriculture Industry Codes, *Journal of Official Statistics*, 3, 375-387.

[106] Rubin, D.B., Stern, H.S. and Vehovar, V. (1995), Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. *Journal of the American Statistical Association*, 90, 822-828.

[107] Sande, G. (1979), Numerical Edit and Imputation, *Paper Presented to the International Association for Statistical Computing*, 42nd Session of the Inrenational Statistical Institute.

[108] Sande, I.G. (1982), Imputation in Survey: Coping With Reality, *The American Statistician*, 36, 145- 152.

[109] Sande, I.G. (1983), Hot-Deck Imputation Procedures, In: *Incomplete Data in Sample Survey, Volume 3*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[110] Sarndal, C.E. (1978), Design-based and Model-based Inference in Survey Sampling, *Scandinevian Journal of Statistics*, 5, 27-52.

[111] Sarndal, C.E. (1980), On π-inverse Weighting versus Best Linear Unbiased Weighting in Probability Sampling, *Biometrika*, 67, 639-650.

[112] Sarndal, C.E. (1996), Efficient Estimators With Simple Variance in Unequal Probability Sampling, *Journal of The American Statistical Association*, 91, 1289-1300.

[113] Sarndal, C.E. and Swensson, B. (1987), A General View of Estimation for Two Phases of Selection with Applications to Two-Phase Sampling and Nonresponse, *International Statistical Review*, 55, 279-294.

[114] Sarndal, C.E., Swensson, B. and Wretman, J.H. (1989), The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total, *Biometrika*, 76, 527-537.

[115] Sarndal, C.R., Swensson, B. and Wretman, J.H. (1992) *Model Assisted Survey Sampling*, John Wiley:New York.

[116] Satin, A. and Shastry, W. (1993), *Survey Sampling: A Non-mathematical Guide*, 2nd eds, Stiatistics Canada: Ottawa, 1993.

[117] Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*, Chapman&Hall:London.

[118] Schaible, W.L. (1983), Estimation of Finite Population Totals from Incomplete Sample Data: Prediction Approach, In: *Incomplete Data in Sample Survey, Volume 3*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[119] Scheaffer, R.L., Mendenhall, W. and Ott, R.L. (1996) *Elementary Survey Sampling, 5th edition*, Duxbury: Boston.

[120] Schenker, N. and Taylor, J.M.G. (1996), Partially Parametric Techniques for Multiple Imputation, *Computational Statistics and Data Analysis*, 22, 425-446.

[121] Schenker, N. and Welsh, A.H. (1988), Asymptotic Results for Multiple Imputation, *The Annals of Statistics*, 16, 1550-1566.

[122] Sedransk, J. and Jinn, J.H. (1992), Secondary Data Analysis When There Are Missing Observations, *Journal of the American Statistical Association*, 87, 952-961.

[123] Sedransk, J., Jinn J.H. and Wang, R. (1991), The Use of Imputed Value in Secondary Data Analysis, *Proceedings of the Seventh Annual Census Bureau Research Confernece*, 483-499.

[124] Sedransk, J. and Smith, P.J. (1988), Inference for Finite Population Quantiles, pp 267-288 In: *Handbook of Statistics 6: Sampling*, Eds: Krishanaiah, P.R. and Rao, C.R., North-holland Publishing:New York.

[125] Sen, A. and Srivastava, M. (1990) *Regression Analysis: Theory, Methods and Applications*, Springer-Verlag:New York.

[126] Shao, J. and Sitter R.R. (1996), Bootstrap for Imputed Survey Data, *Journal of the American Statistical Association*, 91, 1278-1288.

[127] Singh, B. (1983), Bayesian Approach, In: *Incomplete Data in Sample Survey, Volume 2*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[128] Singh, B. and Sedransk, J. (1978), A two-Phase Sampling Design for Estimating the Finite Population Mean when there is Nonresponse, pp 143-155 In: *Survey Sampling and measurement*, ed, Namboodiri N.K., Academic Press:New York.

[129] Skinner, C.J., Holt, D. and Smith, T.M.F. (1989) *Analysis of Complex Surveys*, John Wiley:New York.

[130] Smith, T.M.F (1976), The Foundation of Survey Sampling: A Review, *Journal of the Royal Statistical Society Series A*, 139, 183-204.

[131] Som, R.K. (1996) *Practical Sampling Techniques. 2nd edition, revised and expanded*, Marcel Dekker:New York.

[132] Srinath, K.P. (1971), Multiphase Sampling in Nonresponse Problems, *Journal of the American Statistical Association*, 66, 583-586.

[133] Sudman, S. (1998), Response Effects in Sample Surveys: In *The Encyclopedia* Eds, Armitage, P.A. and Colton, T., pp 3818-3823.

[134] Sukhatme, P.V., Sukhatme, V., Sukhatme, S. and Asok, C. (1984), *Sampling Theory of Surveys with Applications. 3rd edition*, Iowa State University Press.

[135] Sunter, A. (1986), Solutions to the Problem of Unequal Probability Sampling Without Replacement, *International Statistical Review*, 54, 33-50.

[136] Tanner, M.A. (1996) *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 3rd edition*, Springer-Verlag: New York.

[137] Thompson, M.E. (1997) *Theory of Sample Survey*, Chapman&Hall: London.

[138] Thompson, S.K. (1992) *Sampling*, John Wiley: New York.

[139] Thomsen, I. and Siring, E. (1983), On the Causes and Effects of Nonresponse: Norwegian Experience. In:*Incomplete Data in Sample Survey, Volume 3*, Eds: Madow, W.G. and Olkin, I., Acdemic Press, New York.

[140] Thomsen, I. and Tesfu, D. (1988), On the Use of Models in Sampling from Finite Populations, pp 369-396 In: *Handbook of Statistics 6: Sampling*, Eds: Krishanaiah, P.R. and Rao, C.R., North-holland Publishing:New York.

[141] United Nation (1982), *Construction and Use of Sampling Frames from Census Listing*, Ecinomic and Social Commission for Asia and the pacific: Bangkok.

[142] United Nation (1994), *Strategies for Measuring Industrial Structure and Growth*, Department for Economic and Social Information and Policy Analysis:New York.

[143] US Bureau of the Census (1970), *POPSTAN: Evaluation on the Census of Population and Housing*, New York.

[144] Vacek, P.M. and Ahikaga, T. (1980), An Examination of the Nearest Neighbor Rule for Imputing Missing Values, *Proceedings of the Statisitcal Computing Section*, American Statisitcal Association, 326-331.

[145] Valliant, R. (1993), Post-stratification and Conditional Variance Estimation, *Journal of the American Statistical Association*, 88, 89-96.

[146] Wang, R., Sedransk, J. and Jinn, J.H. (1992) Secondary Data Analysis When There are Missing Observations, *Journal of the American Statistical Association*, 87, 952-961.

[147] Wolter, K.M. (1985), *Introduction to variance estimation*, Springer-Verlag: New York.

[148] Yates, F. (1981) *Sampling Methods for Censuses and Surveys. 4th edition*, Macmillian:New York.

[149] Zeger, S.L. and Karim, M.R. (1991), Generalized Linear Models With Random Effects: A Gibbs Sampling Approach, *Journal of the American Statistical Association*, 86, 79-86.